The authors investigate the effect of casino openings on the sales of lottery games in Maryland. This is clearly a policy-relevant research issue and the authors make use of an excellent dataset of sales at each lottery terminal in Maryland. This allows them to construct sales observations at the zip-code-month-level. With some improvements, I think the authors can make a valuable contribution to the literature on optimal lottery design.

Comments (in order of appearance):

1) The first two sections of the article make statements that lack proper citations.

> **Example 1:** "Although there have been analyses of the inter-industry relationships between casinos and lotteries, many of these studies were published at a time when casinos were relatively isolated."
>
> **Example 2:** "Nevertheless, early evidence still suggested that lotteries and casinos were substitutes to a degree."
>
> The presence of a literature review section does not abrogate the requirement of supporting claims with evidence in earlier sections.
>
> Further, existing citations are often out-dated or inappropriate.
>
> **Example 1:** "The most recent and comprehensive analysis of the general relationships among gambling industries is by Walker and Jackson (2008)."
>
> I believe Walker and Jackson (2011) in Contemporary Economic Policy, an article cited by the authors elsewhere, is the appropriate citation.
>
> **Example 2:** "Lotteries received much attention in the literature during the 1980s and early '90s, particularly focused on factors explaining their adoption, their regressive nature, and cross-border purchases.[4]" [*FN4: For a comprehensive discussion of lotteries see Clotfelter and Cook (1991).*]
>
> The authors seem to dismiss that lotteries have received a great deal of attention over the past 15 years, and that a good amount of that attention has been focused on the extent to which lotteries are substitutes or complements, both within the portfolio of games offered by lottery retailers, and with other forms of gambling. In my opinion, these studies are germane to the question at hand and simply referring to Clotfelter and Cook, an article now 23 years old, is not a sufficient presentation. I would suggest the authors at least consider more recent work on the subject, such as: Forrest, Gulley, and Simmons, *Applied Economics*, 2004; Grote and Matheson, *Atlantic Economic Journal*, 2006; Humphries and Perez, *Empirical Economics*, 2012; Trousdale and Dunn, *National Tax Journal*, 2014.

2) The description of lottery and casino gambling in Maryland lacks important information about how lottery revenues are allocated.

> Do proceeds go into a general account that can be used for any purpose? Do proceeds go into an account that can only be used for proscribed purposes? Are the revenues from lottery

treated differently than revenues from casinos? These are important details that influence how we weigh the redistribution of revenue across gambling media.

3) The authors aggregate to the zip-code level and I find their argument for using zip codes over counties to be relatively convincing. I would ask that they address two particular issues.

First, the zip code centroid is never defined. Is it a population centroid or an area centroid? I doubt that zip codes are large enough (except in rural areas) so that the choice of one over the other would greatly affect the calculation of driving time, but it should still be defined.

Second, why zip codes over Census Block Groups or Tracts? The benefit of using either of these definitions is that we have fairly good information about the demographics of communities at Census defined geographies. This leads to my next point…

4) Is there heterogeneity in the response? Do certain types of areas tend to respond more to a decrease in travel time to a casino than others?

Given the concerns about the regressivity of lottery gambling, it would certainly be important to know whether casinos siphon off tax revenue from middle class gamblers. Does this affect how regressive lottery gambling is?

5) Articles in this literature are often silent on what the objective function of state gaming operators. Mason, Steagall and Fabritius (1997) have a wonderful, though often ignored, discussion of this issue that I believe the authors should address in a revision.

Casinos offer much better returns on the gambling dollar. The casino take on a slot machine is 10-15% compared to 50% on most lotto games. Presumably, the gambling consumer walks away better off from this arrangement. Thus, revenue may be down, but utility has increased. And, depending upon who actually goes to casinos, this plays into the distribution argument.

I am not suggesting that the authors "answer" this question, but their discussion should be much expanded to thoughtfully address these issues. I consider the discussion of heterogeneity, redistribution, and the presumed objectives of the lottery operator to be of paramount importance if the authors view this article as a meaningful extension to the report they prepared for the MSL.

6) The empirical specification, because it is presented in seasonal differences, is poorly explained.

I would recommend that the seasonal difference notation be Δ_{12} so that it is clear you are differencing observations 12 units apart in time-space.

The authors should start with a model of sales revenue in levels. Seasonal differencing would eliminate some parameters (zip code and month fixed effects) while redefining others (zip code linear time trends would become zip code fixed effects). Right now, it appears that there are zip code fixed-effects, so these are linear time trends in levels. I am not sure how to interpret the month fixed effects, however. Are the authors saying that there are month

specific year-on-year growth rates? Why would that happen? Given the seasonal differencing, I would get rid of the month fixed effects in your specification.

I think one of the neat things about seasonal differencing is that it allows for observation specific seasonality. If you start with a model in levels that includes zip-code-month fixed effects, these also are differenced out. Thus, you can make an argument that you are allowing for all sorts of unobserved heterogeneity, e.g., "We allow Baltimore to spike differently during tax and holiday seasons than Annapolis." Right now, the myriad benefits of seasonal differencing are largely lost on your audience.

7) Given the stickiness in sales (gambling is addictive, after all), you should consider including the lagged sales as an explanatory variable.

Your estimate is probably close to the long run estimate of the marginal effect, but likely overstates the short-run effect by a large amount. Including lagged sales would lead to endogeneity, but this can be accommodated with a simple IV strategy. These issues should at least be acknowledged and discussed.

8) It seems odd that sales are in levels, rather than logs. Given the enormous skew that large jackpots generate, I would verify that your results are robust to this choice.

9) I am doubtful that any simple specification of distance accurately captures the relationship between proximity to a casino and likelihood of playing.

My prior is that there is a highly non-linear relationship between distance to a casino and lottery sales. Given the large number of observations available, I don't see why the authors can't use a high order polynomial or other flexible form to estimate this relationship between distance and sales. It appears they only considered $1/d$ and $1/d^2$ separately, but not together, or other potential combinations.

10) The authors lack a simple negative test.

Does the current seasonal difference in lottery sales predict a future change in distance to a casino. If so, I would be concerned that the authors are picking up dynamics unrelated to the actual introduction of the casino. A revision should include these results reported.

11) I am concerned by the level of temporal aggregation.

Using monthly observations in some ways artificially inflates the size of your sample. I don't see any mention that the authors clustered their standard errors by zip code. I would also like to see what an estimate using quarterly sales would look like.

12) The seasonal differencing answers the question of what happens to sales one year after a casino opens, but it is a strong assumption that allows us to say that is what the effect is 2 or 3 or 5 years after the casino opens.

This should be noted by the authors and the limitations of the approach acknowledged. They have a long enough sample that they could estimate the model using 24 and 36 month differences.

13) If casinos have only a short-run impact on lottery sales, then the proposed method overestimates the long-run and short-run effects

> Here is a simple example: in year 0, sales are 10 and distance is 10; in year 1, sales are 5 and distance is 5; in year 2, sales are 10 and distance is 5. The long-run effect is zero, but because of the bounce-back in sales, the short-run coefficient on distance would be larger than -1. This returns to the issue of dynamics raised in comment 7, where perhaps lagged distance should also be included in the specification.

14) I have never seen the average of fixed-effects reported, as in Tables 2 and 3, reported as a constant. The average fixed-effect has no meaning unless it is sales weighted.

15) The discussion needs to be greatly expanded.

Referee Report - The effect of casino proximity on lottery sales: Zip code-level evidence from Maryland (NTJ-MS-2014-113)

Summary

The paper develops evidence about the effect of casino openings on retail sales of state lottery products in Maryland. Since both lottery products and casinos generate revenues for state government, the effect of a new casino opening has considerable policy interest. The explanatory variable of interest reflects proximity of retail lottery outlets (aggregated to the ZIP code level) to casinos in Maryland and nearby states. The results of a two way fixed effect model suggest that new casinos openings in Maryland led to a decrease in lottery sales of approximately $44-50 million per year, or 2.7% of annual sales.

Suggestions for Improving the Paper
- Exposition

The writing can be improved. The paper contains a number of very long paragraphs that are difficult to read. For example the paragraph that starts on page 2 is a run-on paragraph that includes both the description of the key result in the paper and the boilerplate "roadmap" paragraph. Figure 4 is unreadable in grey scale.

- Empirical Analysis

Needs significant improvement. As written I find the results in the paper unconvincing.

Aggregation: I understand that aggregating the data to the ZIP code level makes your life easier. Unfortunately, that is not a valid reason for doing this. The dismissal of data dis-aggregated at the retail location is entirely unconvincing: "although we could attempt to model sales at any particular retailer as a function of distance to casinos, such a microlevel analysis would be extremely tedious to perform and interpret. It would also be unlikely to provide advantages over a simpler, more aggregated approach." Tedious? Sorry, but I only care about the quality of your results, not about your personal comfort level with things like GIS software. You claim that the retail-outlet level analysis is "unlikely to provide advantages" over an aggregated analysis. Why? I can think of many reasons why the micro level analysis would be preferred. How about aggregation bias? How about throwing out variation in the data set? How about the fact that keno/monitor lottery products are sold at different retail outlets than lottery tickets and scratch-offs?

Choice of explanatory variable: Your measure of proximity to casinos is based on driving time, not distance. I find this inappropriate. Maryland is an urban state, with bad traffic in the Baltimore and suburban DC counties. " ... we believe that travel time is likely the most operative factor in consumers' behavior in choosing a casino." Drive time is a factor, but it is not clear to me that it is the most important factor. In any event, the drive time variable may or may not reflect that amount of time it

actually takes to get to a certain casino from a given ZIP code. Consider your example on page 10, ZIP 21012 in Arnold. The paper cites a 'typical' drive time to Delaware of 70 minutes and a 64.5 minutes to Perryville. Getting to either involved driving through or around Baltimore. Does this time reflect actual travel time at 4pm on any weekday? I wonder about the drive time from the surburban DC counties, which would involve driving through/around DC. I have no idea what the results mean, because I don't think 'typical' drive times are meaningful in this sort of environment. The presence of traffic means that the variability of actual drive time around this 'typical' drive time is substantial, and systematically different from the 'typical' drive time. At least actual distance in miles does not suffer from this problem.

Then, this drive time variable is transformed by taking the inverse (1/drive time). Why? If proximity is important, then why not include a direct measure of proximity in the regression model? Note that this inverse transformation is not mean preserving, and thus has an unknown effect on the distribution of the drive time variable.

And then, this variable is differenced at a 12 month interval. I have no idea what this transformation does to the distribution of the drive time variable. But here's what occurs to me: in all months when there is no new casino opening that is nearer to each ZIP code than the existing casinos, the value of the differenced variable is zero - so this generates many zeros in this variable. In months around the opening of a new casino that is closer to the ZIP code than an existing casino, this variable is negative (by construction, since the drive time variable is only replaced if the new casino is a closer drive). So this transformation takes an easily interpretable variable (drive time to the nearest casino) and turns it into some mess that takes either a value of zero or a negative value. And this is supposed to represent proximity.

Why not estimate a model with the level of lottery sales at the ZIP code level as the dependent variable, and use a series of month dummy variables to control for seasonal effects. And use the distance from the nearest casino as an explanatory variable. This would appear to be the natural 'baseline' model to get at the question you want to answer. Could it be that this model does not generate any 'interesting' results?

Fixed Effects: The paper estimates a two way fixed effects model, with ZIP code specific intercept shifters and month-year intercept shifters. Because there is unobserved heterogeneity in the data at the ZIP code and month of sample level that needs to be controlled for. You difference the dependent variable, making it year-over-year sales in different lottery products. This should eliminate the ZIP code fixed effects. So putting them in the model invites spurious correlation. Why do it?

And why write out Equation (4) like that? There is no constant in the model you estimate. The actual composite constant term is a_i and tau_t where i and t are subscripts. IT is not an average. And a_i should be zero, since you have differenced that out of the model. Reporting this regression model specification is misleading, and reporting the average constant in the results is wrong.

Interpretation of the results: 'The magnitudes of the coefficients are not immediately obvious because they are in dollar terms, so they do not provide context to determine whether the impact is economically important.' What other terms would the parameter estimate be in to make it easier to determine the economic significance? Utils? Bitcoin? The parameter estimate shows the marginal effect of a ZIP code being one unit closer to a casino (whatever this means - see the comment above) on monthly lottery sales.

'Additionally, a clear interpretation of the coefficients is complicated because we are dealing with a panel of zip codes of very diverse sizes and distances to casinos.' Writing statements like this in a paper erodes the readers confidence that the regression model being estimated is actually understood. The ZIP code fixed effect means that ZIP code size is held constant in the estimates from the regression model. The explanatory variable of interest reflects a (convoluted) measure of distance to a casino.

Tabulated results: The tables include p-values of 0.000. This is impossible. Of course any statistical software will spit out a p-value rounded to zero. And some people will dutifully copy this into the tables. Also, the tables contain parameter estimates like 229.3958. Do you think that the data are precise enough to make numbers four places to the right of the decimal point meaningful?

Why not take into account the actual characteristics of the casinos? Like square footage or number of slot machines.