Improving spatial microsimulation estimates of health outcomes by including geographic indicators of health behaviour: The example of problem gambling

Authors: Markham, F.[1], Young, M.[2], & Doran, B.[1]

1. Australian National University.

2. Southern Cross University

This is a post-peer review 'eprint' of a manuscript that was published in Health and Place in 2017. Full citation details of the final formatted paper are below:

Markham, F., Young, M., & Doran, B. (2017). Improving spatial microsimulation estimates of health outcomes by including geographic indicators of health behaviour: The example of problem gambling. *Health & Place*, *46*, 29–36. https://doi.org/10.1016/j.healthplace.2017.04.008

Abstract

Gambling is an important public health issue, with recent estimates ranking it as the third largest contributor of disability adjusted life years lost to ill-health. However, no studies to date have estimated the spatial distribution of gambling-related harm in small areas on the basis of surveys of problem gambling. This study extends spatial microsimulation approaches to include a spatially-referenced measure of health behaviour as a constraint variable in order to better estimate the spatial distribution of problem gambling. Specifically, this study allocates georeferenced electronic gaming machine expenditure data to small residential areas using a Huff model. This study demonstrates how the incorporation of auxiliary spatial data on health behaviors such as gambling expenditure can improve spatial microsimulation estimates of health outcomes like problem gambling.

Introduction

Background

Problem gambling, characterised by difficulties limiting time and money spent gambling, is a significant and growing public health issue. Harms arising from problem gambling often include financial stress, deteriorated mental and physical health, strained interpersonal relationships, violence and crime. The serious nature of these impacts, combined with their relatively high prevalence in the population, means that problem gambling is in aggregate a serious public health burden. For example, problem gambling has been estimated to be the third-largest contributor to the burden of disability in Victoria, Australia, following major depression and alcohol abuse and dependence (Browne et al., 2016).

Despite its significance as a public health problem, little is currently known about the spatial distribution of problem gambling. Unpublished administrative data on gambling expenditure tends to show highly uneven spatial distributions, suggestive of gambling-related health inequalities. Yet few scholars have specifically examined the spatial distribution of gambling losses. One notable exception is Rintoul *et al.*'s (2013) study, which found that per capita electronic gaming machine (EGM) expenditure was highly concentrated in the most disadvantaged areas of Melbourne. More frequently, the spatial distribution of gambling venues has been mapped and correlated with indicators of deprivation or socioeconomic disadvantage. For example, studies of betting shops in London in 1966 (Newman, 1972) and 2010 (Wardle et al., 2014) show that a historical spatial concentration in more deprived neighbours continues to contemporary times. Similar spatial relationships between EGM venue density and disadvantage have been consistently observed in Australia, Canada, and New Zealand (e.g. Marshall & Baker, 2002; Rush et al., 2007; Wheeler et al., 2006). Moreover, the relationship between venue density and disadvantage may be robust to changes in scale, with modest spatial correlations evident for small geographic zones (with an average of 225 dwellings), as well as for much larger spatial units with populations measured in the tens of thousands (Marshall & Baker, 2001).

The uneven provisioning of gambling venues and gambling expenditure suggests that the prevalence of problem gambling is also likely to be spatially patterned. Yet the degree to which the health burden of problem gambling is spatially uneven is currently unknown.

Put simply, it is unclear if residents of some areas suffer from the adverse impacts of gambling more than others.

A spatial approach to modelling the prevalence of problem gambling is required in order to understand these geographic health inequalities. Beyond an academic imperative to understand the distribution of gambling harms, knowledge of the location of areas of high and low problem gambling prevalence would be useful for a range of practical applications. For example, licensing authorities are typically required to undertake local social impact assessments when new gambling venues are proposed, a task which is difficult to undertake in the absence of local data on the prevalence of problem gambling. Similarly, resources for treatment services ought to be provisioned on the basis of local needs. In short, there are both academic and practical imperatives to understand the spatial distribution of problem gambling.

Yet no studies to date have explicitly sought to estimate the prevalence of problem gambling in small areas. Five notable studies have, however, sought to map the distribution of what Welsh *et al.* term 'debtogenic landscapes' (2014) - urban environments conducive to, or symptomatic of, problem gambling. Taking a combinatory approach, Robitaille and Herjean (2008) mapped the demographic risk factors for problem gambling (i.e. gender, age, income, marital status, income, ethnicity and employment status) and found a spatial correlation between areas of high-risk demographics and the accessibility of gambling venues. Doran and Young (2010) undertook a conceptually similar study, but used index modelling and substituted an index of disadvantage derived using principle components analysis in place of Robitaille and Herjean's separate risk factor layers. This methodology that has since been replicated (Conway, 2015). Rintoul *et al.* (2013) extended this approach, weighting accessibility scores for venues by the volume of EGM expenditure within those venues, rather than following Doran and Young's approach of weighting venues by number of EGMs. The most comprehensive study to date has been that of Wardle *et al.* (2016). This study produced a weighted linear combination of a wide range of risk factors for, and indicators of, problem gambling. They measured not just socio-demographic risk but also the location and utilisation of various mental health services (including problem gambling treatment), the residential location of people utilising homelessness services, and the location of payday-loan outlets and food banks.

The strength of these studies is that they capture the spatial variations of a wide range of gambling-related variables. However, their chief shortcoming is that they are entirely predictive. The outcome variable they produce is a unitless measure of vulnerability, but this index is not calibrated against any empirical data on outcomes *per se*. Consequently, the weights that are assigned to the various elements of vulnerability indices are necessarily arbitrary, with no empirical grounding beyond expert opinion. In effect, they operate in a manner similar to a spatial version of multiple linear regression in which all coefficient values are determined *a priori* by the analyst rather than being estimated from data. At best, the maps produced using this approach provide an educated guess regarding the location and relative prevalence of problem gambling.

This shortcoming is unfortunate given the collection of a large quantity of survey data specifically designed to investigate problem gambling (Williams et al., 2012). The primary limitation of existing surveys that hinders their use in the production of small-area estimates of problem gambling is that they are typically not geocoded (or geocodes are obscured for privacy reasons), so it is difficult to precisely allocate survey responses to residential locations. Even where geocodes are provided, surveys generally do not collect sufficiently spatially-dense data to produce estimates of harm at fine spatial resolutions using regression-based methods such as multilevel modelling (Whitworth et al., 2016).

Other methods such as spatial microsimulation provide an attractive means of producing small area estimates. This paper shows how the strengths of the index modelling approaches discussed above can be combined with well-developed spatial methods to improve small-area estimates. Specifically, spatial microsimulation is used to produce empirically-calibrated small-area estimates of problem gambling that take advantage of spatially-referenced administrative data as well as census data to constrain estimates.

Improving spatial microsimulation estimates of health outcomes with geographic indicators of risk

Spatial microsimulation provides a suite of methods for geographically allocating survey responses to small spatial areas using well-defined spatial data about the small areas to constrain estimates. The purpose is to synthesise a set of geographically-specific study populations, which can then be further analysed in a manner relevant to the study domain and research questions (Lovelace & Dumont, 2016). In typical usage, spatial

microsimulation involves three discrete steps. First, the total counts of persons across different socio-demographic categories are extracted from a population census at the finest possible geographic scale, either as counts of a single census category or as counts from a cross-tabulation of two or more variables. Second, these census-derived totals are harmonised with variables measuring the same construct (e.g. sex, age bracket, etc.) from a survey for which unit record data are available. The outcome variables of interest, which are measured by the survey but not the census, are also identified and included in the unit record data. Third, spatial microsimulation methods are used to allocate survey responses to small areas in a manner that makes the synthesised small-area totals match the census margins as closely as possible. This enables reliable estimates of the outcome variables of interest to be produced at finer geographic scales than those possible using the survey alone.

Spatial microsimulation has been used in this manner to produce small-area estimates of a range of health outcomes. For example, Cataife (2014) combined survey data with census statistics to produce estimates of the prevalence of obesity in tracts spanning just a few city blocks. Similarly, Smith *et al.* (2011) estimated smoking prevalence in Census Area Units in New Zealand, synthesising a national health survey with census data on four socio-demographic variables. These examples share a standard approach to spatial microsimulation in which survey responses are combined with census data without recourse to other sources of spatial information.

However, the reliability of the estimates produced by these methods depends in large part on the ability of census variables to predict the health outcome of interest. In general, the choice of constraint variables is crucial in producing reliable spatial microsimulation based estimates (Smith et al., 2011). In cases where the outcome measure is strongly related to a small number of census variables or their interactions, spatial microsimulation is likely to produce good results. However, for many policy-relevant problems, the outcome of interest is only poorly correlated with census variables. This makes the use of spatial microsimulation less attractive and suggests a need for further, spatially-referenced constraint variables that may not be provided in population censuses.

Environmental risk factors play a role in mediating many health outcomes and provide a likely candidate for providing such additional information. Variables measuring environmental risk factors (e.g. EGM accessibility) have been incorporated into the

problem gambling index models described above (e.g. Conway, 2015; Doran & Young, 2010; Robitaille & Herjean, 2008). In a study aimed at estimating the uptake of gestational diabetes screening in small areas in Ireland, Cullinan *et al.* (2012) provide an example of how auxiliary spatial information on risk can be used to augment typical spatial microsimulation approaches. Because screening uptake is highly dependent on the spatial accessibility of screening facilities, an application of spatial microsimulation to census data alone would have provided geographically questionable results. Therefore, using geocoded hospital register data, the authors converted absolute spatial measures (i.e. individuals' residential latitude and longitude) into a relative spatial measure (i.e. distance to nearest screening centre) and incorporated this as a constraint variable into their model. They were also able to extract other contextual variables such as urban or rural status for each person in the register on the basis of their residential location. These relative-spatial attributes from the register were combined with census data and GIS-calculated data using spatial microsimulation to produce improved small area estimates of screening rates. As this study demonstrated, the inclusion of spatial information above-and-beyond census marginal totals is possible and may indeed be required in some cases to generate sensible spatial microsimulation models.

However, the best predictors of health outcomes are often health-related behaviours. For example, in the case of gambling, socio-demographic variables typically explain around 10% of variance in the problem gambling classification of individuals, while the inclusion of gambling expenditure variables increases variance explained to around 30% (Markham et al., 2016). Geographic indicators of health behaviours are sometimes available and have been included in spatial index models of vulnerability (e.g. Rintoul et al., 2013; Wardle et al., 2016), but rarely in spatial microsimulation studies. We suggest that the inclusion of health behavioural variables in spatial microsimulation analyses is likely to improve the reliability of small area estimates. If surveys provide measures of health behaviours as well as health outcomes, then spatial data relating to health behaviours can provide a crucial link between aggregate collective behaviour at the small area and aggregate health outcomes. This requires the creation of constraint variables for small areas measuring health behaviours that can augment census-derived marginal totals. Census-derived and health-behavioural constraints can then be combined with survey data in spatial microsimulation models.

This solution poses additional problems at the data processing stage. In particular, the transformation of aggregate data relating to health behaviours into categorical constraints for small areas is not always straightforward. We suggest that the answer to these questions is likely to be domain specific. In the case of gambling, the spatial behaviour of consumers is already reasonably well understood (Markham, Doran, et al., 2014), meaning that point-based data on EGM expenditure can be converted to mean per capita expenditure estimates for small areas on the basis of a statistical model. The conversion of population means to numbers of people in different gambling involvement categories can be made on the basis of the distribution of behavioural measures in the survey itself. This prior knowledge can be used as the basis for estimates of mean gambling losses in small areas. Analogous, domain-specific conversations are likely to be possible for other research problems.

Objectives

This study aims to demonstrate the potential for geographical indicators of health behaviours to improve small area estimates derived using spatial microsimulation, with reference to the particular example of estimating problem gambling prevalence. Specifically, this study aims to:

1. compare the explanatory power of individual level models with models including the following predictor variables: a) census variables, b) environmental risk factors, c) health-behavioural measures, and d) a combination of the most important variables across the three categories.
2. compare small-area estimates produced using spatial microsimulation across these three model configurations.

These objectives are pursued in the context of estimating the prevalence of problem gambling in small census areas.

Materials and methods

Setting

This setting for this study is the urban areas of the Northern Territory (NT) of Australia, primarily the towns of Darwin, Katherine and Alice Springs, and their peri-urban hinterlands. At the time of data collection, 88% of EGMs in the NT were located in or adjacent to these three towns, dispersed across 64 licensed gambling venues. The two

largest EGM venues in the study area were the casinos in Alice Springs and Darwin, which together contained more than half of the approximately 2000 EGMs in these towns. The remaining EGMs were distributed among 36 hotels (with approximately 350 EGMs) and 26 clubs (with over 600 EGMs). Clubs are formally not-for-profit community centres, such as sporting or returned servicepersons clubs and were allowed a maximum of 45 EGMs per venue. Hotels or pubs are commercial businesses and were limited to a maximum of 10 EGMs per venue. The EGMs offered by these venues – known as 'poker machines' in the Australian vernacular – were high-intensity machines slot machines, with no minimum spin rate and a maximum bet of $5 per spin, resulting in an average cost of high-intensity gambling of approximately $600 per hour (Productivity Commission, 2010). EGMs can be loaded with up to $1000 at a time. No regulations enforced limit setting by gamblers or breaks in gambling sessions.

Data

The primary data set of interest is a geocoded survey conducted in the urban areas of the NT. Between April and September 2010, a questionnaire was mailed to all 46,263 households in the study area to which Australia Post would deliver unsolicited mail, and a further 2300 questionnaires were hand delivered to peri-urban addresses beyond the range of the postal service. The sample frame was derived from the Australian geocoded national address file (G-NAF), and excluded areas zoned for non-residential uses. Any adult in the household was eligible to participate. The Human Research Ethics Committee of Charles Darwin University granted approval to conduct the study (protocol no. H09048). The questionnaire elicited information on socio-demographics (age, sex, Indigenous status, marital status, and education), gambling behaviour (venues visited, and EGM gambling participation, frequency and session length), and problem gambling (measured using the Problem Gambling Severity Index or PGSI: Ferris & Wynne, 2001). Because the G-NAF was used as a sample frame, all responses could be precisely geocoded to the dwelling level with a 100% match rate. Neighbourhood disadvantage was measured on the basis of residential location, using a census-derived index of economic resources (IER). The IER is produced using a principal components analysis by the Australian Bureau of Statistics (2013) at the Statistical Area 1 (SA1) level of aggregation, a spatial unit with a median population of approximately 400 people. IER values in the study area were discretised into terciles, with the lowest tercile representing areas with the fewest economic resources.

Data to match the survey questions on age, sex, Indigenous status, marital status and education were derived from the 2011 Australian Bureau of Statistics Census of Population and Housing at the SA1 level of aggregation. Age and sex were cross-tabulated to produce separate marginal totals for age brackets (18-39 years, 40-54 years and 55 years or older) for each sex. All other variables were extracted as total counts of single variables for each SA1.

Accessibility to EGM venues is a well-documented environmental risk factor of problem gambling (e.g. Pearce et al., 2008; Welte et al., 2004; Young et al., 2012). An EGM accessibility surface was developed using an unconstrained spatial interaction model by maximising the log-likelihood equation derived by Fotheringham and O'Kelly (1989). EGM venue locations were manually geocoded and a range of attractiveness variables were collected, including: number of EGMs, venue license category, whether the venue was a tourist-oriented inner city bar, proximity to shopping centres, distance to the central business district, and whether the venue had ocean views. Using participants' responses to a question about which EGM venues they visited in the last 30 days, model parameters were estimated that best predicted their reported travel behaviour. The calibrated accessibility model is presented in Equation 1, where: $access_j$ indicates the accessibility score of respondent j; d_{ij} indicates the distance in km between EGM venue i and the home of respondent j; and other indicators represent the attractiveness variables described above.

$$access_j = \sum_i d_{ij}^{-0.83} \cdot size_i^{0.89} \cdot club_i^{0.48} \cdot casino_i^{-0.06} \cdot touristbar_i^{-0.25} \cdot shopcentre_i^{0.20} \cdot \log(dist\_cbd)_i^{0.24} \cdot ocean_i^{0.23} \quad (1)$$

The calibrated parameters of the accessibility model indicate that propensity to visit venues is only weakly related to distance to venue. The gradient of the distance decay curve is relatively flat, with an exponent of -0.83 (95% C.I.: -0.81, -0.85) suggesting that accessibility impacts on visitation behaviour at a regional scale rather than a highly localised scale (W. Hansen, 1959). Venue size is also crucial to accessibility, with a venue with 45 EGMs contributing 3.8 (95% C.I.: 3.5, 4.2) time more to accessibility than a venue with 10 EGMs. Clubs also contributed more to EGM accessibility than hotels, while venues that were located close to supermarkets or that had ocean views were also

more accessible. Accessibility scores were calculated for each respondent and discretised into terciles.

The health behaviour of interest was gambling involvement. Involvement was measured for local areas using per capita gambling expenditure. Mean per capita gambling expenditure in each SA1 was estimated from administrative data on EGM expenditure for individual venues during the survey period provided by the NT Department of Justice. These authoritative data are considered complete and reliable because they are generated from a computerised centralised monitoring system. The monitoring system collects real-time transaction data from each EGM in the NT, an arrangement designed to prevent EGMs being used to facilitate organised crime (Australian Institute for Gambling Research, 1999). A previously published Huff model that was calibrated on the same data set was used to allocate expenditure at EGM venues to local areas (Markham, Doran, et al., 2014; Markham, Young, et al., 2014), producing an estimate of mean per capita expenditure for each SA1. The mean expenditure estimates were used as a basis for calculating the number of persons with differing gambling involvement levels in each SA1. Specifically, minutes spent gambling on EGMs in the last 30 days was calculated for each respondent on the basis of their survey responses. Expenditure was derived via time because the survey instrument did not contain questions about money lost gambling. Time spent gambling was converted to dollars spent gambling using the population-weighted mean EGM expenditure velocity, which was calculated to be $2.35 per minute. Survey estimated dollars spent per month for individuals and mean per capita EGM expenditure were combined at the SA1 level. A bivariate regression analysis was conducted on these SA1-level data to estimate the relationship between mean per capita expenditure and the percentage of residents with high EGM gambling involvement (defined as expenditure of $300 in the last 30 days) or no EGM gambling involvement in the last 30 days. The remaining individuals spending $1-$299 in the last thirty days classified as low involvement (see Figure 1), and was calculated as the remaining population in each SA1 after the other two groups had been accounted for. These relationships were used to estimate the number of people in each gambling involvement category in each SA1.

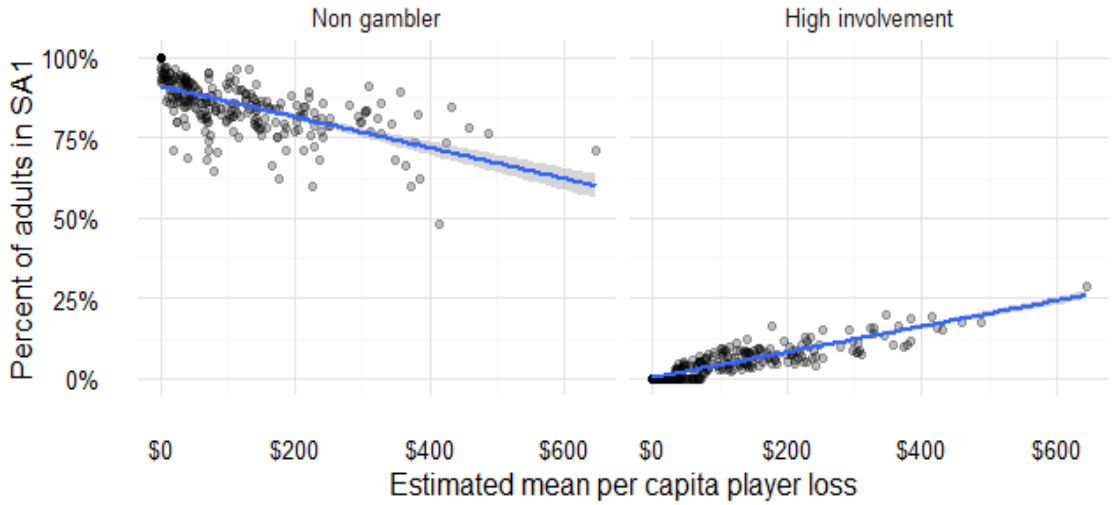

Figure 1: Estimated relationship between Huff model derived mean per capita EGM expenditure and percentage of respondents with survey-derived gambling involvement. Units of analysis were SA1s. Regression lines are weighted by the number of survey respondents in each SA1. R^2 for non-gamblers and high involvement categories were 0.41 and 0.78 respectively.

Ultimately, two data sets were assembled, each covering the same data items at different scales of aggregation, one for individuals primarily derived from the survey, and another of total counts aggregated to the SA1 level. A summary of these variables and their data sources is provided in Table 1.

Table 1: Summary of variables used in the spatial microsimulation analysis and their data sources

Variable	Individual-level source	SA1-level source
Age and sex	Survey	Census
Indigenous status	Survey	Census
Education	Survey	Census
Marital status	Survey	Census
Neighbourhood disadvantage	ABS IER for SA1 of survey respondent's residence	ABS IER, discretised into terciles.
Accessibility	Calculated for each survey respondent and discretised into terciles.	Calculated for each dwelling in the sample frame, with the proportion of dwellings in each SA1 in each discrete category calculated, with the number of persons in each category imputed from these proportions.
EGM gambling involvement	Survey	Calculated from a Huff model which allocates administrative data on expenditure in EGM venues to SA1s. SA1 mean per capita expenditure is converted into numbers of persons in

		each involvement category using the regression estimates presented in Figure 1. Categories were “No EGM participation”, “Low EGM participation” and “High EGM participation”.
Problem gambling	Survey	Not applicable, this is the outcome variable.

Statistical analysis

A two-phase approach was taken to the statistical analysis. In the first phase, the power of three sets of variables (socio-demographic, environmental and involvement) to explain problem gambling risk were explored. All models were limited to four predictor variables as previous research has suggested that the inclusion of too many constraints reduces the performance of spatial microsimulations (Tanton & Edwards, 2012). Four logistic regression models were fit to predict whether or not an individual would meet the conventional classification of a problem gambler (a score eight or more on the PGSI). The first model contained socio-demographic variables only. The second model contained only two environmental risk factors, accessibility and neighbourhood

Table 2: Multiple logistic regression coefficients and indices of model fit for four different sets of variables predicting problem gambling among individuals

	Socio-demographic model		Environmental risk factor model		Gambling involvement model		Combined model	
	O.R.	95% C.I.	O.R.	95% C.I.	O.R.	95% C.I.	O.R.	95% C.I.
Intercept	0.02	0.01, 0.04	0.03	0.02, 0.04	0.01	0.01, 0.01	0.01	0.00, 0.01
Female, aged 18-39 years	1.00						1.00	
Female, aged 40-54 years	1.69	0.93, 3.19					1.47	0.78, 2.87
Female, aged 55 years or older	1.15	0.58, 2.29					0.66	0.33, 1.35
Male, aged 18-39 years	**6.29**	**3.46, 11.90**					**5.31**	**2.78, 10.51**
Male, aged 40-54 years	1.95	0.98, 3.93					1.60	0.77, 3.33
Male, aged 55 years or older	1.43	0.73, 2.84					0.84	0.42, 1.73
Married or in a de facto marriage	**0.51**	**0.36, 0.72**					**0.55**	**0.38, 0.80**
Indigenous	**4.20**	**2.46, 6.89**					**3.13**	**1.70, 5.54**
Attained school-level qualifications	1.00							
Attained technical qualifications	**0.48**	**0.28, 0.80**						
Attained university qualifications	**0.53**	**0.36, 0.77**						
Low accessibility tercile			1.00					
Medium accessibility tercile			1.23	0.81, 1.90				
High accessibility tercile			1.39	0.93, 2.11				
Low I.E.R. tercile			1.00					
Medium I.E.R. tercile			**0.62**	**0.43, 0.90**				
High I.E.R. tercile			**0.42**	**0.27, 0.64**				
Non E.G.M. gambler					1.00		1.00	
Spent between $1 and $300 on E.G.M.s in last 30 days					**6.43**	**3.95, 10.35**	**6.14**	**3.62, 10.20**
Spent $300 or more on E.G.M.s in last 30 days					**40.55**	**27.46, 60.66**	**43.52**	**28.42, 67.60**
A.I.C.	1260		1422		1111		985	
Pseudo R^2	0.14		0.02		0.23		0.33	

Notes: O.R. = odds ratio, C.I. = confidence interval, I.E.R. = index of economic resources, E.G.M. = electronic gaming machine, A.I.C. = Akaike's Information Criterion. All reported odds ratios are adjusted for other variables in the model. Bold type indicates odds ratios whose 95% confidence intervals do not contain 1.0.

disadvantage. The third model included only a single measure of health behaviour, 'gambling involvement', defined on the basis of EGM expenditure. The final model included the four predictor variables across all categories that best fit the data drawn from a total of seven

possible predictor variables. Akaike's Information Criterion (AIC) and McFadden's pseudo-R^2 were reported as relative measures of model fit. Multicollinearity among predictor variables was unusually low, with generalised variance inflation factors calculated as less than 1.5 in all cases.

Finally, four spatial microsimulation models were run using the same four combinations of predictor variables as the logistic regression analysis. Combinatorial optimisation was used to allocate individual survey respondents to SA1s. An implementation of simulated annealing was used to minimise total absolute error when synthesising the population of each SA1, with the maximum number of iterations set to 1000. The *sms* package in *R* was used to undertake this computation (Kavroudakis, 2015). The prevalence of problem gambling was calculated among these synthesised populations at the SA1 level. SA1-level estimates from the four models were compared cartographically and formally tested for statistical correlations.

Results

In total, 7049 people completed the survey, resulting in a response rate of 14.5%. The prevalence of problem gambling in the entire sample was 2.1% (95% C.I. 1.8%, 2.5%). Respondents were more likely to be female (61.9%, 95% C.I. 60.7%, 63.0%), and aged 55 or older (36.6%, 95% C.I 35.5%, 37.8%) than the general population in the study area. Mean imputed 30-day EGM gambling expenditure, calculated from reported time spent gambling, was $114 among the sample (SD=$470). The distribution of expenditure was highly right skewed as is typical for this kind of data, with 86.7% of the sample not participating in EGM gambling at all.

The logistic regression models of problem gambling show that gambling involvement is the single best predictor of problem gambling among individuals (see Table 2). The model including only gambling involvement explained 23% of the variance in problem gambling classification. In contrast, socio-demographic variables and environmental risk variables explained just 14% and 2% of variance respectively. Combining the gambling involvement variable with selected socio-demographic variables produced the best fitting model, which explained 33% of the variance in problem gambling classifications. The four variables to include in this combined model were selected on the basis of AIC.

All variables except accessibility were significantly correlated with problem gambling risk. In particular, men aged between 18 and 39 were 5–6 times more likely to be problem gamblers than women of that same age. Indigenous people were 3 or 4 times more likely to report problem gambling. Those who were married or in facto relationships, those who had completed post-school education, and those who lived in wealthier areas were half as likely to report problem gambling. Finally, those who were imputed to have spent $300 or more on EGM gambling in the last 30 days were 40 times more likely to report problem gambling than those who didn't gamble on EGMs in the same period.

The problem gambling prevalence estimated by the four spatial microsimulation were rather similar when analysed collectively for the entire study area. The socio-demographic model estimated problem gambling prevalence at 2.3%, the environmental risk factor model estimated 2.0%, the gambling involvement model estimated a prevalence of 2.1%, while the combined model estimated a population problem gambling prevalence of 2.2%. However, the spatial patterning of problem gambling changed substantially depending on model configuration. As Table 3 demonstrates, the prevalence of problem gambling at the SA1 level was only significantly correlated for the combined model and the socio-demographic model. Even in this case, the correlations were weak. This divergence is evident when problem gambling prevalence estimates are mapped spatially. Figure 2 shows the spatial distribution of problem gambling prevalence in north Darwin, an important region in the study area. In some areas, model predictions are relatively consistent, for example southern Tiwi, where prevalence estimates ranged from between 1.6% (Panel A) and 2.0% (Panels D). In contrast, prevalence estimates varied substantially between models in other areas like part of northern Leanyer, with estimates ranging from 0.7% (Panel B) to 4.1% (Panel A). In general, both Table 3 and Figure 2 demonstrate that the environmental risk factor model produces results that are dissimilar to those produced by the other three models.

Table 3: Correlation matrix of problem gambling prevalence estimates for SA1s produced using four spatial microsimulation models

	Socio-demographic model	Environmental risk factor model	Gambling involvement model	Combined model
Socio-demographic model	1.0			
Environmental risk factor model	-0.02	1.0		
Gambling involvement model	-0.02	0.03	1.0	
Combined model	**0.25**	-0.10	-0.01	1.0

Notes: Pearson's correlation coefficients are reported. Bold type indicates correlations that are significant at the $p < 0.05$ level.

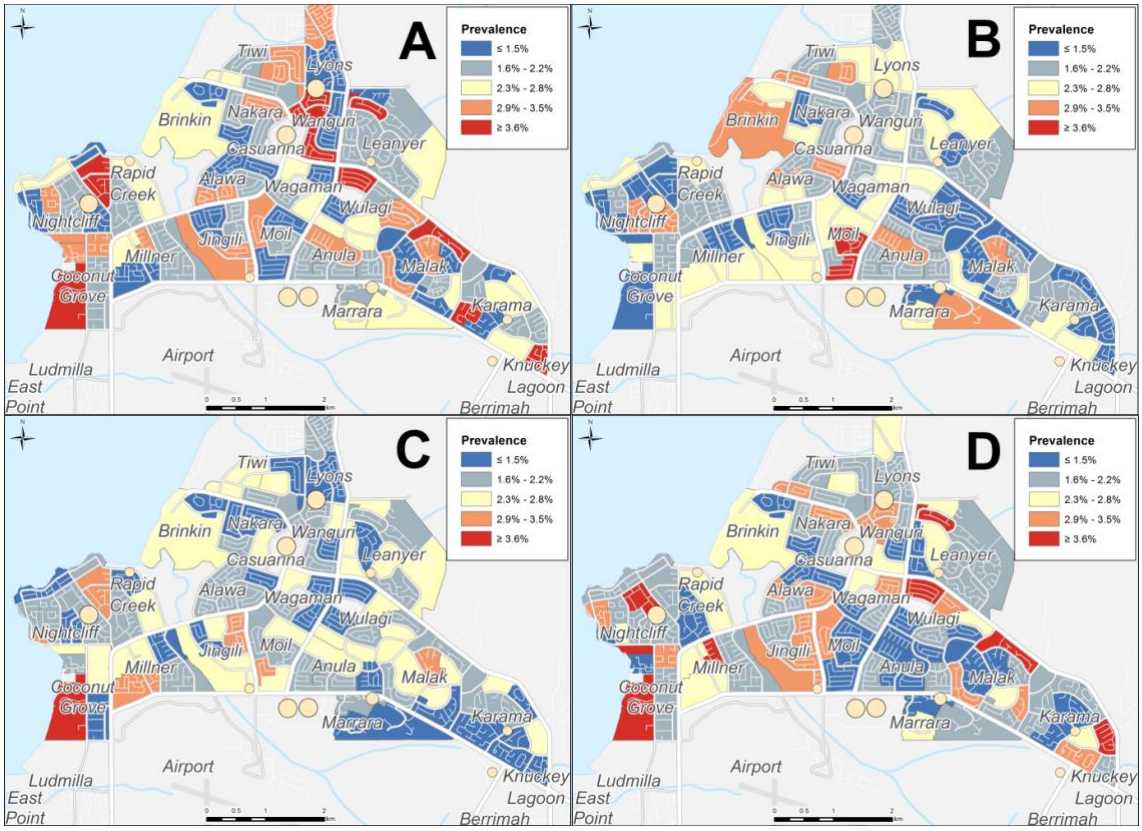

Figure 2: Maps of estimated prevalence of problem gambling in part of the study area generated from four spatial microsimulation models. Panel A shows predictions from the socio-demographic model. Panel B shows predictions from the environmental risk model. Panel C shows predictions from the gambling involvement model. Panel D shows predictions from the combined model.

Discussion

Interpretation of key results

This study has demonstrated how problem gambling prevalence estimates for small areas can be empirically-derived by combining survey data with census and health behavioural data. It has two key findings. First, extending the approach of Cullinan *et al.* (2012), it has shown how spatial microsimulation can incorporate constraints beyond socio-demographic variables and include measures of environmental risk factors and health behaviours. Logistic regression analysis suggests that the combination of health-behavioural measures and socio-demographic variables has the greatest explanatory power in terms of predicting health outcomes. In the case of problem gambling, the addition of a behavioural measure (gambling involvement) to a socio-demographic model increased the pseudo R^2 from 0.14 to 0.33. This demonstrates the potential afforded by the incorporation of health behavioural measures. Second, this study has shown that spatial microsimulation studies are highly sensitive to model specification. Different model specifications can produce markedly different spatial patterns of the outcome variable of interest.

This study demonstrates the utility of marshalling administrative data on EGM expenditure in venues to predict gambling involvement rates in small geographic areas. Figure 1 demonstrated a remarkably high correlation between mean per capita expenditure in an SA1, estimated using a Huff model, and survey-derived estimates gambling involvement for those same small area ($R^2 = 0.78$). Put plainly, this means that with the aid of a Huff model and venue-level EGM expenditure data, the number of people in a particular location gambling at a high intensity can be predicted with a remarkable degree of accuracy. Such a finding provides further evidence for the validity of Rose and Day's (1990) total consumption theory to the study of EGM expenditure (cf. M. Hansen & Rossow, 2008; Lund, 2008; Markham, Young, et al., 2014).

It is perhaps to be expected that health behavioural measures improve the predictive power of models of health outcomes. After all, in most cases health behaviours have greater causal proximity to outcomes than do environmental risk factors or socio-demographic variables. It is surprising, therefore, that health-behavioural variables have rarely been incorporated into spatial microsimulation studies as constraints. One reason for this absence is the usual omission of health-behavioural measures from census data. This study contributes to the literature by demonstrating how point-structured administrative data can be converted into constraints for small areas and then used for the purpose of spatial microsimulation.

The specification dependency among spatial microsimulation model results warrants further investigation. What is especially surprising is that the combined model, which incorporates the measure of gambling involvement, the single most explanatory variable, produced small-area estimates that were uncorrelated with those produced on the basis of gambling involvement alone. This might be explained by the optimisation goal of minimising total absolute error in a context of low multicollinearity. The combined model specification contained three socio-demographic variables but only one gambling involvement variable. Because total absolute error counts deviations from each constraint category equally regardless of the variable's explanatory power, the inclusion of a greater number of socio-demographic variables might 'weight' the analysis toward the constructs represented by the greatest number of constraint variables. A similar observation has been made in the case of cluster analysis (Hair et al., 2009), although the difference in the case of spatial microsimulation is that this problem is likely to be mitigated – not exacerbated – by the use of multicollinear predictor variables.

The feature selection weighting effect warrants future research into both its impact on results and into methods for mitigating it. One potential mitigation measure may be to duplicate a constraint that is under-weighted, thereby doubling its contribution to calculating total absolute error.[1] This approach might be generalised through the inclusion of arbitrary feature weights in total absolute error calculations in the combinatorial optimisation process. In this case, weights could be defined on the basis of principal components analysis or other methods of feature reduction. These suggested modifications to spatial microsimulation methods warrant future research, but are beyond the scope of this study. Until methods are developed for dealing with the feature selection weighting effect, analysts using spatial microsimulation should specify their models with a great deal of care and on the basis of theoretical concerns as well as goodness-of-fit indices.

The specification dependence exhibited in these results appear to be of more concern than those discussed previously in the peer-reviewed literature (e.g. Smith et al., 2009). Such variation deriving from sensitivity to model specification is not accounted for in recent methods developed to quantifying uncertainty in spatial microsimulation estimates (Nagle et al., 2014; Whitworth et al., 2016). While model specification uncertainty is by no means unique to spatial microsimulation, the results suggest that the problem may be especially acute when using this

[1] The authors are indebted to an anonymous reviewer for this suggestion.

method. Model averaging provides one promising avenue by which model specification uncertainty may be quantified. The implication of sensitivity to model specification is that users of spatial microsimulation need to exercise great caution in ensuring that results are robust to variations in model configuration. This is especially important when, as in this case study, there is no 'gold standard' data against which external validation can take place. We believe that this finding is likely to be generalizable across geographic locations and problem domains in cases when the outcome variable of interest is only moderately correlated with the predictor variables, as is very frequently the case (e.g. Anderson, 2007; Whitworth et al., 2016). The improvement of model fit through the addition of health-behavioural measures is likely to be especially valuable in such situations.

Conclusions

The purpose of this paper was to explore the benefits of including health-behavioural variables in spatial microsimulation studies of health outcomes, with specific reference to problem gambling and gambling involvement. The study has made four contributions to the literature. First, it found that including health behavioural variables in a spatial microsimulation analysis was not only viable, but dramatically improved the explanatory power of related statistical models. This approach to incorporating auxiliary information should be encouraged in future applications of these methods. Second, the study demonstrated the accuracy of predicting gambling involvement in small areas on the basis of EGM expenditure data reported for individual gambling venues. Specifically, it found that the proportion of residents in small areas reporting high-gambling involvement in a survey could be accurately predicted on the basis of administrative data regarding gambling expenditure. Third, the inclusion of a health behavioural variable also demonstrated that spatial microsimulation results are dependent on model specification to an extent not generally appreciated in the literature. In cases where external validation against gold-standard data is not possible, sensitivity to model specification should be explicitly investigated. In cases with high sensitivity to model specification, results should be interpreted with caution. Future research might usefully develop and evaluate methods for assigning arbitrary weights to constraints when in the combinatory optimisation process. Finally, this study has provided four sets of empirically-calibrated estimates of problem gambling prevalence in small areas. It demonstrates that a great degree of spatial inequality exists in the prevalence of problem gambling, an inequality that is not only of concern in its own right, but also plays a role in furthering disparities among other economic and health outcomes. Such inequalities demand urgent policy attention.

References

Anderson, B. (2007). *Creating Small Area Income Estimates for England: Spatial microsimulation modelling* (Chimera Working Paper No. 2007–07). University of Essex. http://opendepot.org/166 archived at http://www.webcitation.org/6mPHIxL5g

Australian Bureau of Statistics. (2013). *Technical Paper: Socio-Economic Indexes for Areas (SEIFA), Australia*. http://www.abs.gov.au/ausstats/abs@.nsf/mf/2033.0.55.001 archived at http://www.webcitation.org/6mQ75xZ0Q

Australian Institute for Gambling Research. (1999). *Australian Gambling Comparative History and Analysis*. Victorian Casino and Gaming Authority. https://assets.justice.vic.gov.au/vcglr/resources/bb81f943-d854-40de-8bab-b09d8bbd610f/australiangamblingcomparativehistory.pdf archived at http://www.webcitation.org/6cNpqq7Xn

Browne, M., Langham, E., Rawat, V., Greer, N., Li, E., Rose, J., Rockloff, M., Donaldson, P., Thorne, H., Goodwin, B., Bryden, G., & Best, T. (2016). *Assessing gambling-related harm in Victoria: A public health perspective*. Victorian Responsible Gambling Foundation. https://www.responsiblegambling.vic.gov.au/__data/assets/pdf_file/0007/28465/Browne_assessing_gambling-related_harm_in_Vic_Apr_2016-REPLACEMENT2.pdf archived at http://www.webcitation.org/6pEx9apZZ

Cataife, G. (2014). Small area estimation of obesity prevalence and dietary patterns: A model applied to Rio de Janeiro city, Brazil. *Health & Place*, *26*, 47–52. https://doi.org/10.1016/j.healthplace.2013.12.004

Conway, M. (2015). Vulnerability modeling of casinos in the United States: A case study of Philadelphia. *Applied Geography*, *63*, 21–32. https://doi.org/10.1016/j.apgeog.2015.05.015

Cullinan, J., Gillespie, P., Owens, L., & Dunne, F. (2012). Accessibility and screening uptake rates for gestational diabetes mellitus in Ireland. *Health & Place*, *18*(2), 339–348. https://doi.org/10.1016/j.healthplace.2011.11.001

Doran, B., & Young, M. (2010). Predicting the spatial distribution of gambling vulnerability: An application of gravity modeling using ABS Mesh Blocks. *Applied Geography*, *30*(1), 141–152. https://doi.org/10.1016/j.apgeog.2009.04.002

Ferris, J., & Wynne, H. (2001). *The Canadian Problem Gambling Index: User Manual*. Canadian Centre on Substance Abuse. http://www.ccsa.ca/2003%20and%20earlier%20CCSA%20Documents/ccsa-009381-2001.pdf archived at http://www.webcitation.org/67tGlKq7P

Fotheringham, S., & O'Kelly, M. E. (1989). *Spatial Interaction Models: Formulations and Applications*. Kluwer Academic Publishers.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate Data Analysis* (7 edition). Pearson.

Hansen, M., & Rossow, I. (2008). Adolescent gambling and problem gambling: Does the total consumption model apply? *Journal of Gambling Studies*, *24*(2), 135–149. https://doi.org/10.1007/s10899-007-9082-4

Hansen, W. (1959). How Accessibility Shapes Land Use. *Journal of the American Planning Association*, *25*, 73–76. https://doi.org/10.1080/01944365908978307

Kavroudakis, D. (2015). sms: An R Package for the Construction of Microdata for Geographical Analysis. *Journal of Statistical Software*, *68*(2), 1–23. https://doi.org/10.18637/jss.v068.i02

Lovelace, R., & Dumont, M. (2016). *Spatial Microsimulation with R*. CRC Press.

Lund, I. (2008). The population mean and the proportion of frequent gamblers: Is the theory of total consumption valid for gambling? *Journal of Gambling Studies*, *24*(2), 247–256. https://doi.org/10.1007/s10899-007-9081-5

Markham, F., Doran, B., & Young, M. (2014). Estimating gambling venue catchments for impact assessment using a calibrated gravity model. *International Journal of Geographical Information Science*, *28*(2), 326–342. https://doi.org/10.1080/13658816.2013.838770

Markham, F., Young, M., & Doran, B. (2014). Gambling expenditure predicts harm: Evidence from a venue-level study. *Addiction, 109*(9), 1509–1516. https://doi.org/10.1111/add.12595

Markham, F., Young, M., & Doran, B. (2016). The relationship between player losses and gambling-related harm: Evidence from nationally representative cross-sectional surveys in four countries. *Addiction, 111*(2), 320–330. https://doi.org/10.1111/add.13178

Marshall, D., & Baker, R. G. V. (2001). Clubs, spades, diamonds and disadvantage: The geography of electronic gaming machines in Melbourne. *Australian Geographical Studies*, *39*(1), 17–33. https://doi.org/10.1111/1467-8470.00127

Marshall, D., & Baker, R. G. V. (2002). The evolving market structures of gambling: Case studies modelling the socioeconomic assignment of gaming machines in Melbourne and Sydney, Australia. *Journal of Gambling Studies*, *18*(3), 273–291. https://doi.org/10.1023/A:1016847305942

Nagle, N. N., Buttenfield, B. P., Leyk, S., & Spielman, S. (2014). Dasymetric modeling and uncertainty. *Annals of the Association of American Geographers*, *104*(1), 80–95. https://doi.org/10.1080/00045608.2013.843439

Newman, O. (1972). *Gambling: Hazard and Reward.* Athlone Press. http://prism.ucalgary.ca//handle/1880/41340

Pearce, J., Mason, K., Hiscock, R., & Day, P. (2008). A national study of neighborhood access to gambling opportunities and individual gambling behaviour. *Journal of Epidemiology and Community Health, 62,* 862–868. https://doi.org/10.1136/jech.2007.068114

Productivity Commission. (2010). *Gambling* (Report No. 50). Productivity Commission.

Rintoul, A. C., Livingstone, C., Mellor, A. P., & Jolley, D. (2013). Modelling vulnerability to gambling related harm: How disadvantage predicts gambling losses. *Addiction Research & Theory, 21*(4), 329–338. https://doi.org/10.3109/16066359.2012.727507

Robitaille, E., & Herjean, P. (2008). An analysis of the accessibility of video lottery terminals: The case of Montreal. *International Journal of Health Geographics, 7*(2), 1–15. https://doi.org/10.1186/1476-072X-7-2

Rose, G., & Day, S. (1990). The population mean predicts the number of deviant individuals. *BMJ: British Medical Journal, 301*(6759), 1031–1034.

Rush, B., Veldhuizen, S., & Adlaf, E. (2007). Mapping the prevalence of problem gambling and its association with treatment accessibility and proximity to gambling venues. *Journal of Gambling Issues, 20,* 193–214.

Smith, D. M., Clarke, G. P., & Harland, K. (2009). Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A, 41*(5), 1251–1268. https://doi.org/10.1068/a4147

Smith, D. M., Pearce, J. R., & Harland, K. (2011). Can a deterministic spatial microsimulation model provide reliable small-area estimates of health behaviours? An example of smoking prevalence in New Zealand. *Health & Place, 17*(2), 618–624. https://doi.org/10.1016/j.healthplace.2011.01.001

Tanton, R., & Edwards, K. L. (2012). Limits of Static Spatial Microsimulation Models. In R. Tanton & K. Edwards (Eds.), *Spatial Microsimulation: A Reference Guide for Users* (pp. 161–168). Springer Netherlands. https://doi.org/10.1007/978-94-007-4623-7_10

Wardle, H., Astbury, G., Thurstain-Goodwin, M., & Parker, S. (2016). *Exploring Area-Based Vulnerability to Gambling-Related Harm: Developing the Gambling-Related Harm Risk Index*. City of Westminster. http://transact.westminster.gov.uk/docstores/publications_store/licensing/final_phase2_exploring_area_based_vulnerability_to_gambling_related_harm.pdf archived at http://www.webcitation.org/6mMoxbRE8

Wardle, H., Keily, R., Astbury, G., & Reith, G. (2014). 'Risky places?': Mapping gambling machine density and socio-economic deprivation. *Journal of Gambling Studies*, *30*(1), 201–212. https://doi.org/10.1007/s10899-012-9349-2

Welsh, M., Jones, R., Pykett, J., & Whitehead, M. (2014). The "problem gambler" and socio-spatial vulnerability. In F. Gobet & M. Schiller (Eds.), *Problem Gambling: Cognition, Prevention and Treatment* (pp. 156–187). Palgrave MacMillan.

Welte, J. W., Wieczorek, W. F., Barnes, G. M., Tidwell, M.-C. O., & Hoffman, J. H. (2004). The relationship of ecological and geographic factors to gambling behavior and pathology. *Journal of Gambling Studies*, *20*(4), 405–423. https://doi.org/10.1007/s10899-004-4582-y

Wheeler, B. W., Rigby, J. E., & Huriwai, T. (2006). Pokies and poverty: Problem gambling risk factor geography in New Zealand. *Health & Place*, *12*(1), 86–96. https://doi.org/10.1016/j.healthplace.2004.10.011

Whitworth, A., Carter, E., Ballas, D., & Moon, G. (2016). Estimating uncertainty in spatial microsimulation approaches to small area estimation: A new approach to solving an

old problem. *Computers, Environment and Urban Systems*, *63*, 50–57. https://doi.org/10.1016/j.compenvurbsys.2016.06.004

Williams, R. J., Volberg, R. A., & Stevens, R. M. G. (2012). *The Population Prevalence of Problem Gambling: Methodological Influences, Standardized Rates, Jurisdictional Differences, and Worldwide Trends* [Technical Report]. Ontario Problem Gambling Research Centre. http://www.uleth.ca/dspace/handle/10133/3068 archived at http://www.webcitation.org/6OK7CeYsW

Young, M., Markham, F., & Doran, B. (2012). Too close to home? The relationships between residential distance to venue and gambling outcomes. *International Gambling Studies*, *12*(2), 257–273. https://doi.org/10.1080/14459795.2012.664159