



Gerogiannis, G., Tranmer, M., Lee, D. and Valente, T. (2022) A Bayesian spatio-network model for multiple adolescent adverse health behaviours. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(2), pp. 271-287.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

This is the peer reviewed version of the following article Gerogiannis, G., Tranmer, M., Lee, D. and Valente, T. (2022) A Bayesian spatio-network model for multiple adolescent adverse health behaviours. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(2), pp. 271-287, which has been published in final form at <https://doi.org/10.1111/rssc.12531>. This article may be used for non-commercial purposes in accordance with [Wiley Terms and Conditions for Self-Archiving](#).

<https://eprints.gla.ac.uk/254084/>

Deposited on: 6 October 2021

# A Bayesian spatio-network model for multiple adolescent adverse health behaviours

George Gerogiannis, Mark Tranmer, Duncan Lee

*University of Glasgow, UK.*

Thomas Valente

*University of South California, USA.*

**Summary.** The use of alcohol, tobacco, and marijuana among adolescents are major public health concerns, and a number of epidemiological studies have been conducted to understand the drivers of these individual health behaviours. However, there is no literature that jointly models these health behaviours with the aim of understanding the relative importances of individual factors, friendship effects, and spatial effects in determining the prevalence of alcohol, tobacco, and marijuana use among adolescents. To address this gap in the literature we propose a novel multivariate spatio-network model for jointly modelling all three of these behaviours, with inference conducted in a Bayesian setting using Markov chain Monte Carlo simulation. The model is motivated by survey data from 5 schools in Los Angeles, California, and the results indicate the important roles that individual factors and friendship networks play in driving the uptake of these health behaviours.

## 1. Introduction

The consumption of alcohol, cigarettes, and marijuana are some of the largest public health concerns around the world. There is a strong consensus in the scientific community that alcohol consumption can cause various forms of cancer, with the National Toxicology Program (NTP) of the US Department of Health and Human Services listing the consumption of alcohol as a known human carcinogen based on evidence from studies in humans ([National Toxicology Program, 2016](#)). Similarly, the epidemiological literature on the negative effects of cigarettes has an incredibly long history ([Doll and Hill, 1950, 1954](#)), and smoking causes cancer, heart disease, stroke, lung diseases, diabetes, and chronic obstructive pulmonary disease (COPD) ([United States Surgeon General, 2014](#)). In contrast, the study of the health effects of marijuana is a relatively new field of research, and a review of studies suggest that marijuana users are significantly more likely than non-users to develop long-lasting mental health disorders such as schizophrenia ([Volkow et al., 2016](#)).

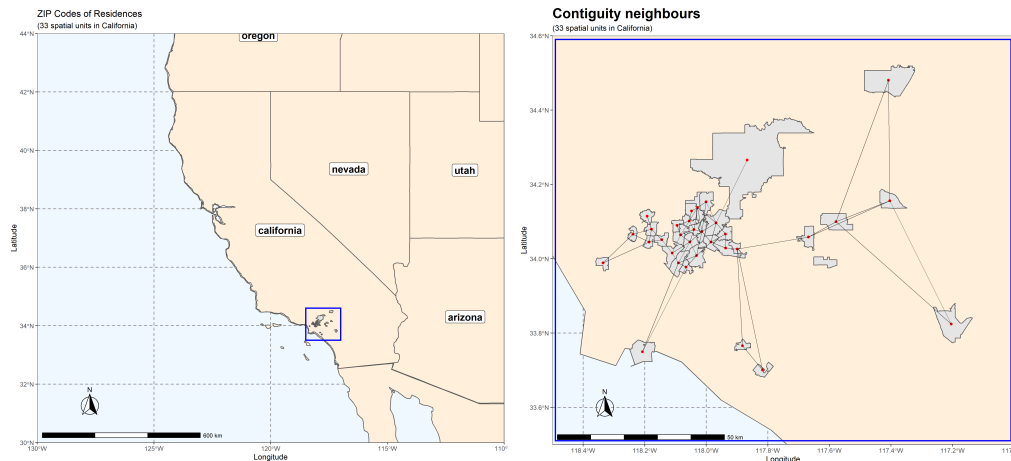
Thus it is of interest to understand the factors that drive people to partake in these inter-related adverse health behaviours, and a number of different factors have been identified. One of the most salient and consistent predictors of an individual's behaviours is peer influence ([Leung et al., 2014](#)), and many theories have been put forward to explain this mechanism. For example, [Bandura and Walters \(1977\)](#) propose a *social learning theory*, in which individuals acquire new behaviours through the observation and imitation of their peers. Individual factors have also been strongly associated with adverse health behaviours, such as educational attainment ([Gilman et al., 2008](#)) and stressful life

events (Simantov et al., 2000). Finally, spatial effects have also been observed (Pearce et al., 2009 and Lee and Lawson, 2016), with rates of adverse health behaviours being higher in some neighbourhoods than others. This then begs the natural question as to which of these factors has the greatest influence on adolescent health behaviours, which is the gap in the literature that this paper addresses. The identification of the drivers governing adolescent health behaviours is crucial in a public health context, because such information can be used in the design of intervention strategies. There are many examples of such interventions and policies aimed at reducing an individual's likelihood of partaking in adverse health behaviours (Lantz et al., 2000), many of which can be classified as neighbourhood (Lewit et al., 1997) or social network (Starkey et al., 2009; Valente, 2012; Brown et al., 2019) strategies.

The modelling approaches used to quantify the effects of these drivers on adverse health behaviours are typically based around Generalized linear models (GLMs) and extensions thereof. For example, Castillo et al. (2017) used a GLM to study whether regional targeting of interventions that aim to reduce the frequency as well as volume of drinking may be effective, while Fujimoto and Valente (2015) sought to provide insight into how adolescent health behaviour is predicated. In contrast, multiple membership multiple classification (MMMC, Browne et al., 2001) approaches extend GLMs by explicitly modelling network effects via random effects representing peer influence in the model. Most notably, Tranmer et al. (2014) and Lorant and Tranmer (2019) use a univariate class of this model for investigating adolescent behaviours, and focus on estimating the relative share of variation of risky and protective health behaviours at different levels of the population structure: individuals, friendship networks, and schools.

These existing studies are limited however, in that they have not fully accounted for the combined effects of individual factors, network effects and spatial location, thus making them unable to discern the relative importance of these factors. Additionally, they typically consider only a single health behaviour, and thus are not able to quantify the similarities in these behaviours and their drivers. For example, Castillo et al. (2017) ignores the spatial dependence that may be present in the data, while Lorant and Tranmer (2019) only consider one health behaviour at a time. This study thus proposes a novel spatio-network modelling approach for inter-related adolescent health behaviours, which for the first time enables the relative importance of individual factors, friendship effects and spatial effects on multiple health behaviours to be quantified. The statistical novelty in our approach is the fusion of network models (Browne et al., 2001) with spatial areal unit models (Leroux et al., 2000), and we provide software via an R package to allow others to implement our methods on their own data. The software in fact provides a suite of models that allow the 3 components (individual factors, friendship network effects and spatial effects) to be included or excluded, allowing a direct comparison of how well each of these factors explain adolescent health behaviours.

This work is organized as follows: Section 2 describes the study that motivates the methodological development, while Section 3 outlines the novel Bayesian spatial-network models proposed. Section 4 presents the application of the models to the motivating study data, while Section 5 provides a concluding discussion.



**Fig. 1.** Maps of California (left) and the spatial configuration of the Zip Codes (right). In the latter the lines denote the neighbour relationships between two Zip Codes assumed when fitting the model.

## 2. Motivating study

The data in this study come from the Social Networking Survey (Valente et al., 2013), which is based in Los Angeles, California (see left panel of Figure 1). The study surveyed 1,068 10<sup>th</sup> grade adolescents (15-16 year olds) from this area who belong to one of five schools in the region, with between 158 and 284 adolescents belonging to each school. Further details of the study are available from Valente et al. (2013).

### 2.1. Health behaviour data

The health behaviour data are binary (no / yes) self-reported measures of alcohol, cigarettes, and marijuana consumption for each 10<sup>th</sup> grade student in 2010. More specifically, whether an individual has: (i) *consumed at least one drink of alcohol in the past 30 days*; (ii) *smoked at least one cigarette in the past 30 days*; and (iii) *ever tried marijuana*. The numbers responding no / yes for each adverse health behaviour are summarised in Table 1, where the results are presented separately for each school. The table shows that over all the schools the highest prevalence was for drinking alcohol (29%), with marijuana use (27%) being the least common adverse health behaviour. There is greater variation in the prevalence of adverse health behaviours between schools, with school 2 having an average prevalence over all three responses of 37% compared with 19% for school 4. Finally, the table shows that there were no missing values in these adverse health behaviour responses.

### 2.2. Covariate data

In addition to data on the school that each adolescent attends, the survey also provides data on possible individual-level factors that may affect the prevalence of each

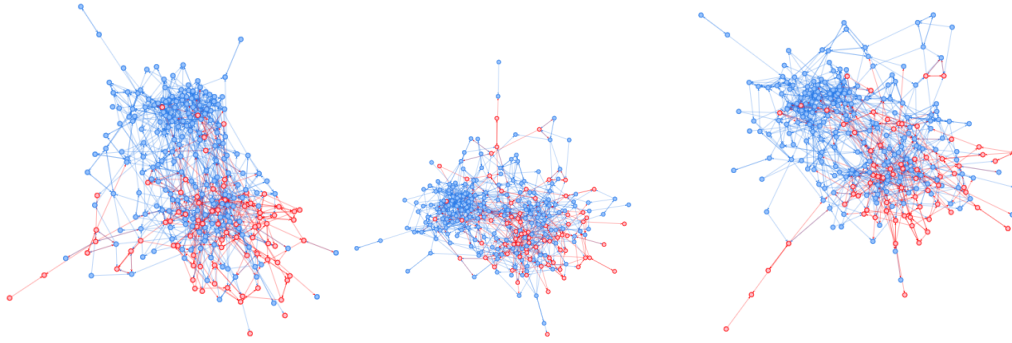
**Table 1.** Summary of the response and covariate data split across the 5 schools.

	School 1 <i>n</i> = 284	School 2 <i>n</i> = 185	School 3 <i>n</i> = 158	School 4 <i>n</i> = 256	School 5 <i>n</i> = 185	All schools <i>n</i> = 1068
<b>Responses</b>						
Alcohol consumption (no/yes)	203 / 81	119 / 66	98 / 60	215 / 41	120 / 65	755 / 313
Cigarette consumption (no/yes)	208 / 76	115 / 70	109 / 49	197 / 59	133 / 52	762 / 306
Marijuana consumption (no/yes)	211 / 73	117 / 68	109 / 49	211 / 45	133 / 52	781 / 287
<b>Gender</b>						
Female	157	84	84	124	108	557
Male	127	101	74	132	77	511
<b>Exam grades</b>						
Mostly A's	48	17	10	38	17	130
Mostly A's and B's	68	35	42	77	53	275
Mostly B's	15	14	12	18	9	68
Mostly B's and C's	79	50	54	61	55	299
Mostly C's or lower	74	69	40	62	51	296

adverse health behaviour. Following an exploratory model building exercise using simple Bernoulli logistic regression models, two covariates exhibited significant effects for all 3 health behaviours and are hence included in our analysis. These covariates are summarised in Table 1, and the first is the gender of the individual which is a well known factor that influences an individual's health behaviour (Institute of Medicine et al., 2001). The table shows there are more females (52%) than males (48%) in the survey, although the difference is not large. The second covariate we consider is an individual's educational attainment, because existing studies (Escobedo and Peddicord, 1996; Hu et al., 2006) have shown this can affect adverse health behaviours. The survey contained a categorical variable denoting the modal grade achieved in each individual's previous year of exams, and the table shows that overall 56% of adolescents achieved mostly B's and C's or lower in their exams, while 12% obtained mostly A's.

### 2.3. Friendship network structure

The friendship structure between the adolescents is captured in the survey by two questions: (i) *Please think of your seven best friends in 10th grade*; and (ii) *Are there other people in the 10th grade who you consider a close friend?* For the latter question each individual was allowed to nominate up to 12 friends, which when combined with the first question resulted in up to 19 nominations in total. The friendship network structure for adolescents in school 1 is displayed in Figure 2 as a directed graph, and similar figures for the remaining schools are not shown for brevity. In the figure each vertex represents an individual, while a line from individual *A* to individual *B* denotes that *A* nominated *B* as a friend. Note this is a directed graph, because if *A* nominates *B* as a friend it does not necessarily follow that *B* also nominates *A*. The three panels of Figure 2 relate to the three adverse health behaviours, and in each case a red vertex is an individual who does partake in the adverse health behaviour while a blue vertex denotes somebody who does not. Visually the figure shows that blue vertices are predominantly connected to blue vertices, while red vertices are mainly connected to other red vertices. This appears to be the case for all three adverse health behaviours, and suggests that adolescents are more likely to partake in these behaviours if they are friends with somebody else who also partakes in the same behaviour.



**Fig. 2.** The directed friendship network structure present in school 1. The individuals (vertices) are coloured by their alcohol (left), tobacco (middle), and marijuana (right) response - red for “yes” and blue for “no”. A directed connection (line) between 2 vertices (individuals) denotes that one nominated the other as a friend.

To assess the presence of such friendship network effects more formally, the empirical probability of engaging in a specific health behaviour given that the individual has nominated at least one other individual that engages in that same health behaviour was computed. In the case of consuming at least one drink of alcohol in the past 30 days these probabilities were 0.74, 0.62, 0.68, 0.60, and 0.8 across the 5 schools. In the case of smoking at least one cigarette in the past 30 days these probabilities were 0.70, 0.57, 0.53, 0.68, and 0.79 across the 5 schools. Finally, in the case of ever trying marijuana these probabilities were 0.73, 0.63, 0.65, 0.71, and 0.75 across the 5 schools. These conditional probabilities are much higher than the marginal probabilities of partaking in each adverse health behaviour, which across all 5 schools are: alcohol - 0.29, tobacco - 0.29, marijuana - 0.27. This suggests the presence of friendship (peer) network effects in the data, which we account for in our proposed model outlined in Section 3.

#### 2.4. Spatial Zip Code structure

The adolescents surveyed collectively reside in  $S = 33$  non-overlapping administrative areas known as Zip Codes, which contain very unequal numbers of survey responders. For example, the two Zip Codes with the largest number of adolescents surveyed are El Monte (91732 - 378 individuals) and South El Monte (91733 - 271 individuals), while there are 19 instances in which only 1 adolescent surveyed resides in the Zip Code. The spatial configuration of the  $S = 33$  Zip Codes is displayed in the right panel of Figure 1, which shows that while most of the Zip Codes are grouped together in the middle of the region, there are a small number of isolated Zip Codes that are not close to the remaining ones.

The spatial closeness between each pair of Zip Codes is encoded in the model described in the next section by a binary neighbourhood matrix denoted  $\mathbf{A}_{33 \times 33}$ , where the  $ij$ th element  $a_{ij} = 1$  if Zip Codes  $(i, j)$  share a common border and  $a_{ij} = 0$  otherwise (and  $a_{ii} = 0$  for all  $i$ ). This *border sharing* specification is the most commonly used neighbourhood matrix in spatial areal unit modelling (see for example Bivand et al., 2013 and Jack et al., 2019), because of its sparsity and simplicity of construction (e.g.

it does not have a tuning parameter as the  $k$ -nearest neighbours rule does). However, Zip Codes that are isolated share no neighbours under this definition, which means the conditional autoregressive prior outlined in the next section for capturing the spatial correlation has improper full conditional distributions for these Zip Codes. Thus we make the commonly used adjustment to  $\mathbf{A}$  for each isolated Zip Code  $i$  to rectify this problem, which is to make them a neighbour of the Zip Code  $j$  that is geographically closest (e.g. set  $a_{ij} = a_{ji} = 1$ ). The final neighbourhood structure assumed when fitting the model is displayed by the connecting lines in the right panel of Figure 1, which shows that under this specification the Zip Codes comprise a single connected graph.

We assessed in an exploratory manner whether there are likely to be spatial effects in the data, i.e. whether adolescents in different Zip Codes have differing propensities for partaking in adverse health behaviours. We do this by computing the empirical probability of engaging in a specific health behaviour given that the individual is from a certain Zip Code. However, as previously discussed the distribution of individuals to the 33 Zip Codes is highly skewed, with a minimum, 1st quartile, median, 3rd quartile, and maximum of 1, 1, 1, 6, and 378 individuals respectively. Thus for a meaningful comparison we only consider the 5 Zip Codes containing the most individuals, which are Temple City (91780 - 31 individuals), El Monte (91731 - 154 individuals), Rosemead (91770 - 167 individuals), South El Monte (91733 - 271 individuals), and El Monte, (91732 - 378 individuals). For these Zip Codes the observed probabilities of consuming at least one drink of alcohol in the past 30 days are 0.16, 0.27, 0.15, 0.33, and 0.28 respectively. The corresponding probabilities of smoking at least one cigarette in the past 30 days are 0.13, 0.36, 0.17, 0.33, and 0.29 respectively, while for marijuana the probabilities are 0.16, 0.27, 0.15, 0.33, and 0.28. As these empirical probabilities show some variation by Zip Code, spatial effects are a plausible component to include in the model.

### 3. Methodology

We propose a novel spatio-network Bayesian hierarchical model for multiple adverse health behaviours, which simultaneously estimates the effects of covariate factors, friendship influence and spatial location.

#### 3.1. Data likelihood model

As the adverse health behaviours are binary no/yes outcomes, the data likelihood model has a Bernoulli logistic regression form and is given by

$$Y_{i_s r} \sim \text{Bernoulli}(\theta_{i_s r}) \quad \text{for } i = 1, \dots, N_s, \quad s = 1, \dots, S, \quad r = 1, \dots, R,$$

$$\ln \left( \frac{\theta_{i_s r}}{1 - \theta_{i_s r}} \right) = \mathbf{x}_{i_s}^\top \boldsymbol{\beta}_r + \phi_{sr} + \sum_{j \in \text{net}(i_s)} w_{i_s j} u_{jr} + w_{i_s}^* u_r^*. \quad (1)$$

Here  $Y_{i_s r}$  denotes the binary ( $Y_{i_s r} = 1$  - yes;  $Y_{i_s r} = 0$  - no) adverse health behaviour for the  $r$ th response for the  $i$ th individual who lives in the  $s$ th spatial unit, where  $r = 1, \dots, R (= 3)$ ,  $s = 1, \dots, S (= 33)$  and  $i = 1, \dots, N_s$ . The probability that individual

$i$  from spatial unit  $s$  partakes in adverse health behaviour  $r$  is denoted by  $\theta_{i_s r}$ , which is modelled on the logit scale by three separate components. The first is a  $p \times 1$  vector of covariates  $\mathbf{x}_{i_s}$ , which is accompanied by a  $p \times 1$  vector of fixed effect regression parameters  $\beta_r$  that vary by health behaviour  $r$ . The prior for these fixed effect parameters is given by  $\beta_r \sim N(\mathbf{0}, 100000\mathbf{I})$  independently for each outcome  $r$ , where  $\mathbf{I}$  is the  $p \times p$  identity matrix. This specification is chosen to be weakly informative, thus allowing the parameter estimates to be largely informed by the data. The other two components in the systematic part of the model are a friendship network effect and a spatial effect, and these two different levels in the model are described below. The friendship network effect captures correlations between the three health behaviours and between individuals (friends), while the spatial effect capture correlations between neighbouring Zip Codes.

### 3.2. Friendship network effects

Friendship network effects are accounted for in the model by the  $\sum_{j \in \text{net}(i_s)} w_{i_s j} u_{jr} + w_{i_s}^* u_r^*$  component in (1), where  $\text{net}(i_s)$  denotes the set of individuals that the  $i$ th individual in the  $s$ th spatial unit nominates as friends. The latter part  $u_r^*$  is an *isolation effect* for health behaviour  $r$ , which is an effect for individuals who don't nominate any friends. This is achieved by setting  $w_{i_s}^* = 1$  if individual  $i_s$  nominates no friends and  $w_{i_s}^* = 0$  otherwise, and if  $w_{i_s}^* = 1$  then clearly  $\sum_{j \in \text{net}(i_s)} w_{i_s j} u_{jr} = 0$  as  $\text{net}(i_s)$  is the empty set. The  $\sum_{j \in \text{net}(i_s)} w_{i_s j} u_{jr}$  component is known as a multiple membership model, and was proposed by Browne et al. (2001). The friendship network structure for each individual is discussed in Section 2.3 and displayed visually in Figure 2, but for the purposes of the model is encoded into an  $n \times n$  friendship network matrix  $\mathbf{W} = (w_{i_s j})$ , where  $n = \sum_{s=1}^S N_s$  denotes the number of individuals in the survey. The values of the entries in this friendship network matrix are given by

$$w_{i_s j} = \begin{cases} 1/|\text{net}(i_s)| & \text{if individual } i \text{ in spatial unit } s \text{ nominates individual } j \text{ as a friend} \\ 0 & \text{Otherwise} \end{cases}, \quad (2)$$

where  $|\text{net}(i_s)|$  is the cardinality of the set  $\text{net}(i_s)$ . Thus, the only non-zero entries in this matrix relate to friendships that one individual has with another individual. Note, this matrix is not necessarily symmetric because it represents a directed rather than an undirected graph. The values of the non-zero elements in this matrix are the reciprocal of the number of friends each individual nominates, which ensures that the matrix is row standardised (each row sums to 1).

Then for individual  $i$  in spatial unit  $s$  the friendship network component of the model contributes the term  $\sum_{j \in \text{net}(i_s)} w_{i_s j} u_{jr}$  to the linear predictor. Here  $u_{jr}$  is a random effect representing the peer effect that person  $j$  has on their friends in terms of partaking in health behaviour  $r$ . Thus  $\sum_{j \in \text{net}(i_s)} w_{i_s j} u_{jr}$  simply represents the average (mean) effect that the friends of individual  $i$  in spatial unit  $s$  have on that individual in terms of partaking in health behaviour  $r$ . We allow these friendship random effects  $\mathbf{u}_j = (u_{1j}, \dots, u_{Rj})_{R \times 1}$  to be correlated across the different responses, because we believe that if an individual encourages others to drink alcohol for example then they may also



encourage others to smoke cigarettes or marijuana as well. Thus we specify the following multivariate normal prior distribution for each individual’s friendship random effects  $\mathbf{u}_j$ :

$$\begin{aligned}\mathbf{u}_j &\sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma}), \\ \mathbf{u}^* &\sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma}), \\ \boldsymbol{\Sigma} &\sim \text{Inverse-Wishart}(4, \mathbf{I}).\end{aligned}\tag{3}$$

Between health behaviour correlation is allowed for each  $\mathbf{u}_j$  and  $\mathbf{u}^*$  via the  $R \times R$  covariance matrix  $\boldsymbol{\Sigma}$ , which is assigned a weakly informative inverse Wishart prior distribution, which allows the estimation to mainly be informed by the data. We assume that  $\mathbf{u}_j$  and  $\mathbf{u}^*$  share the same covariance matrix  $\boldsymbol{\Sigma}$  for convenience.

### 3.3. Spatial effects

Section 2.4 suggests there may be spatial effects in the data, which we model using spatially correlated random effects. These random effects are assigned a conditional autoregressive (CAR) prior distribution, which is the most common approach to modelling spatial correlation in areal unit data (see for example [Banerjee et al., 2004](#)). As our adverse health behaviour response is multivariate, a multivariate CAR model could be adopted to model both spatial and between health behaviour correlations. Multivariate CAR type models are an active research area, and numerous different approaches have been proposed including [Gelfand and Vounatsou \(2003\)](#), [Jin et al. \(2007\)](#), [Martinez-Beneito \(2013\)](#) and [MacNab \(2016\)](#). However, in this paper we model the correlations between the 3 adverse health behaviours via the friendship network component of the model as described above, and thus modelling these correlations a second time may cause parameter identifiability issues in the model. In fact, as we show in the next section, the friendship network effects are a much more important driver of adverse health behaviours than the spatial effect, so the between behaviour correlations are more prominently captured in that component of the model.

Therefore instead we specify independent CAR models for each adverse health behaviour  $r$ , which are specified as a prior distribution for a vector of spatial random effects  $\boldsymbol{\phi}_r = (\phi_{1r}, \dots, \phi_{Sr})$  for the  $S = 33$  Zip Codes. Spatial correlation is induced into these random effects through the spatial neighbourhood matrix  $\mathbf{A}$  described in Section 2.4, which is defined by the commonly used *border sharing* rule. We note here that as the spatial correlation structure is defined by  $\mathbf{A}$ , all inferences about this part of the model are conditional on the choice of  $\mathbf{A}$ . Following a comparative study of different CAR priors by [Lee \(2011\)](#), we use the CAR prior proposed by [Leroux et al. \(2000\)](#) due to its consistent superior performance. This prior has the joint distribution  $\boldsymbol{\phi}_r \sim \text{N}(\mathbf{0}, \tau_r^2[\gamma_r(\text{diag}(\mathbf{A}\mathbf{1}) - \mathbf{A}) + (1 - \gamma_r)\mathbf{I}]^{-1})$  for each response  $r$ , where  $\mathbf{1}$  is an  $S \times 1$  vectors of ones,  $\mathbf{I}$  is the  $S \times S$  identity matrix and  $\text{diag}(\mathbf{A}\mathbf{1})$  is a diagonal matrix with diagonal elements obtained by the matrix product  $\mathbf{A}\mathbf{1}$ . Thus the joint distribution of the spatial random effects for all three health behaviours  $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3)_{3S \times 1}$  is a zero-mean multivariate normal distribution, whose covariance matrix is block diagonal with three  $S \times S$  blocks given by  $\tau_r^2[\gamma_r(\text{diag}(\mathbf{A}\mathbf{1}) - \mathbf{A}) + (1 - \gamma_r)\mathbf{I}]^{-1}$  for  $r = 1, 2, 3$ .

The precision matrix for the spatial random effects relating to health behaviour  $r$  is given by  $\gamma_r(\text{diag}(\mathbf{A}\mathbf{1}) - \mathbf{A}) + (1 - \gamma_r)\mathbf{I}$  and is a weighted average of correlated ( $\text{diag}(\mathbf{A}\mathbf{1}) - \mathbf{A}$ ) and independent ( $\mathbf{I}$ ) components, where the amount of spatial dependence is controlled by  $\gamma_r$ . However, the fact that this prior captures spatial correlation is much easier to see from its univariate full conditional form, which is given, together with the hyperpriors, for response  $r$  by

$$\begin{aligned} \phi_{sr} | \phi_{-sr} &\sim \text{N} \left( \frac{\gamma_r \sum_{j \neq s} a_{sj} \phi_{jr}}{\gamma_r \sum_{j \neq s} a_{sj} + 1 - \gamma_r}, \frac{\tau_r^2}{\gamma_r \sum_{j \neq s} a_{sj} + 1 - \gamma_r} \right), \\ \tau_r &\sim \text{Half-Normal}(10000), \\ \gamma_r &\sim \text{Uniform}(0, 1). \end{aligned} \quad (4)$$

Here  $\phi_{-sr}$  denotes the vector of  $S - 1$  spatial random effects for outcome  $r$  excluding  $\phi_{sr}$ . In this prior the spatial dependence parameter  $\gamma_r$  is assigned a non-informative prior on the unit interval, and if  $\gamma_r = 1$  the model simplifies to the intrinsic CAR prior for strong spatial correlation proposed by Besag et al. (1991) because the conditional expectation is the mean of the random effects in neighbouring areas. In contrast, if  $\gamma_r = 0$  it is trivial to see that the random effects are independent. Finally, a weakly informative (large variance) half normal prior centred on zero is specified for the spatial standard deviation  $\tau_r$  as suggested by Gelman (2006), which again lets the data play the dominant role in the estimation of its value.

### 3.4. Inference and software

Posterior inference from the model was obtained using Markov chain Monte Carlo (MCMC) simulation, including both Gibbs sampling and Metropolis-Hastings steps. Software to implement the MCMC algorithm was written in C++ using the R package Rcpp (Eddelbuettel and François, 2011), and is then made user friendly for future users by providing an R (R Core Team, 2013) wrapper function. The software is freely available to download from <https://github.com/GNG3/modelSoftwareMVBSN>.

The model described above, as well as simplifications of it, are fitted to the study data outlined in Section 2, and the results are presented in the next section. In all cases inference is based on two parallel Markov chains each run for 400,000 iterations, 200,000 of which were discarded as burn-in and the 200,000 post-burn-in iterations were thinned by 20 on the basis of RAM limitations and creating samples which are less correlated. Convergence of the Markov chains was assessed by examining trace plots of the posterior samples for a selection of parameters, the Geweke diagnostic (Geweke, 1992), and the Gelman-Rubin statistic (Gelman and Rubin, 1992), and in all cases the samples appear to have converged. Therefore final inference is based on 20,000 MCMC samples for each parameter, 10,000 from each chain.

## 4. Results from the study

We fit 8 different models to the survey data, which allows us to examine the relative importance of covariate effects, friendship network effects and spatial effects in explaining

**Table 2.** Summary and overall fit of the 8 models.

Model	Covariates	Space	Network	DIC	$p_D$
$\mathcal{M}1$	-	-	-	3,821	3.0
$\mathcal{M}2$	✓	-	-	3,463	30.1
$\mathcal{M}3$	-	✓	-	3,798	22.1
$\mathcal{M}4$	-	-	✓	2,904	427.5
$\mathcal{M}5$	✓	✓	-	3,463	29.9
$\mathcal{M}6$	✓	-	✓	2,891	408.6
$\mathcal{M}7$	-	✓	✓	2,907	417.7
$\mathcal{M}8$	✓	✓	✓	2,883	413.1

an adolescents' propensity to partake in adverse health behaviours. These 8 models are denoted  $\mathcal{M}1$  to  $\mathcal{M}8$  and contain all possible combinations of the three different model components, ranging from  $\mathcal{M}1$  which only contains an intercept term through to  $\mathcal{M}8$  which is the full model given by (1). A summary of the components included in each model is given in Table 2 for ease of reference. The covariates used in these models include the categorical variables gender, exam grades and school, which are summarised in Section 2.2.

#### 4.1. Model comparison

The overall fit of each model is summarised in Table 2, which displays the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002) and the effective number of independent parameters ( $p_D$ ). A comparison of the single component models  $\mathcal{M}2$ ,  $\mathcal{M}3$ , and  $\mathcal{M}4$  to the null (intercept only) model shows that the inclusion of the friendship network component leads to the greatest reduction in DIC compared to the intercept only model, as the DIC goes from 3,821 to 2,904, a reduction of 917. The sole inclusion of covariates have just under half this impact with a DIC reduction of 358, while the sole inclusion of a spatial component leads to a DIC reduction of just 23. Adding in the covariates ( $\mathcal{M}6$ ) and then additionally the spatial component ( $\mathcal{M}8$ ) to the friendship network model improves the fit to the data but only marginally, with the DIC reductions compared to the network only model ( $\mathcal{M}4$ ) being only 13 and 21 respectively. Thus, while the full model with all three components has the lowest DIC value, the impact of adding in the covariates and the spatial components are small once the friendship network effects are included.

The results in Table 2 show that the effective number of independent parameters  $p_D$  went down for  $\mathcal{M}8$  compared to  $\mathcal{M}4$  and  $\mathcal{M}7$ , despite the former model being the most complex in terms of its parameterisation. The reason for this is that in the full model  $\mathcal{M}8$  the variation in the data is jointly modelled by all three components, where as in  $\mathcal{M}4$  for example only the network component is included. Thus in  $\mathcal{M}8$  the network component is having to model less of the variation in the data compared to in  $\mathcal{M}4$ , due to the covariates and to a much lesser degree the spatial effect modelling some of this variation. This results in a reduction in the effective number of independent parameters for the network component in model  $\mathcal{M}8$  compared to  $\mathcal{M}4$  due to a reduction in the variation in the random effects  $\{\mathbf{u}_j\}$ , which thus causes the reduced  $p_D$ . The remainder of this section present the results relating to the full model  $\mathcal{M}8$ , so that the effects of

**Table 3.** Estimates and 95% credible intervals for the between health behaviour correlations.

Adverse health behaviours	Estimated correlation
$\rho_{12} = \Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$ - alcohol and cigarettes	0.956 (0.808, 0.996)
$\rho_{13} = \Sigma_{13}/\sqrt{\Sigma_{11}\Sigma_{33}}$ - alcohol and marijuana	0.955 (0.792, 0.997)
$\rho_{23} = \Sigma_{23}/\sqrt{\Sigma_{22}\Sigma_{33}}$ - cigarettes and marijuana	0.975 (0.890, 0.998)

all three components can be observed.

#### 4.2. Model fit

In order to confirm that the model fits the data adequately we simulate 1,000 trivariate samples  $\{\tilde{\mathbf{y}}^{(1)}, \dots, \tilde{\mathbf{y}}^{(1000)}\}$  from the posterior predictive distribution  $f(\tilde{\mathbf{y}}|\mathbf{y})$ , where  $\mathbf{y}$  denotes the observed data. As both  $(\tilde{\mathbf{y}}^{(j)}, \mathbf{y})$  are binary this posterior predictive check involves computing the probability that the observed data matches the simulated data generated from the posterior predictive distribution. Averaging over all individuals  $i$ , spatial units  $s$ , health behaviour  $r$  and posterior predictive samples  $j$ , the posterior predictive probability  $\mathbb{P}(\tilde{y}_{i_s r} = y_{i_s r}) = 0.68$ , suggesting that the model fits the data relatively well as it generates simulated data that are similar to the real data. The corresponding health behaviour specific values are  $\mathbb{P}(\tilde{y}_{i_s 1} = y_{i_s 1}) = 0.67$ ,  $\mathbb{P}(\tilde{y}_{i_s 2} = y_{i_s 2}) = 0.67$ , and  $\mathbb{P}(\tilde{y}_{i_s 3} = y_{i_s 3}) = 0.70$ , suggesting that the model fits the marijuana response slightly better than the other two.

Additionally, Table 3 provides the posterior means and 95% credible intervals for the between health behaviour correlations from the friendship network component of the model, which allows us to examine the appropriateness of modelling all three health behaviours jointly. These correlations are captured in  $\Sigma$ , and for example the correlation between alcohol and cigarettes is computed by  $\rho_{12} = \Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$ . The table shows that the correlations are very high and close to one for each pair of adverse health behaviours, with posterior means ranging between 0.955 (alcohol and marijuana) and 0.975 (cigarettes and marijuana). These strong correlations thus support the use of a joint modelling approach for our adverse health behaviours. Finally, the posterior samples of these network correlation parameters yield  $\mathbb{P}(\rho_{23} > \rho_{12} \cap \rho_{23} > \rho_{13}) = 0.629$ , suggesting that the correlation between the cigarette and marijuana responses is likely to be greater than both the correlations between the alcohol and cigarette responses and the alcohol and marijuana responses.

#### 4.3. Covariate effects

Table 4 displays the estimated covariate effects (posterior means) and 95% credible intervals for each adverse health behaviour, and all results are presented as odds ratios relative to the baseline level of the factor (the first one in the table denoted by a “-”). The table shows that in comparison to females, the baseline level, males had a significantly reduced odds of consuming alcohol in the past 30 days, with an estimated odds ratio of 0.57. In contrast, the 95% credible intervals for the male covariate relating to the cigarette and marijuana responses show no statistically significant gender effect.

In contrast, the effect of exam performance is much more consistent than that of gender, with decreasing exam performance being significantly associated with higher

**Table 4.** Summary of the covariate effects as odds ratios and selected other parameters from model  $\mathcal{M}8$ .

	Alcohol	Cigarettes	Marijuana
<b>Covariates</b>			
Female	-	-	-
Male	0.57 (0.39, 0.83)	1.31 (0.86, 1.99)	0.91 (0.56, 1.46)
A's	-	-	-
A's and B's	2.49 (1.10, 6.12)	1.41 (0.61, 3.34)	2.48 (0.86, 7.82)
B's	2.66 (0.91, 7.85)	0.68 (0.19, 2.21)	2.35 (0.59, 9.55)
B's and C's	4.16 (1.86, 10.01)	2.60 (1.16, 6.07)	5.73 (2.09, 17.04)
C's or lower	8.05 (3.61, 19.74)	5.46 (2.47, 12.46)	14.66 (5.47, 45.05)
<b>Space</b>			
School 1	-	-	-
School 2	0.96 (0.35, 2.41)	1.12 (0.35, 3.38)	1.07 (0.28, 3.79)
School 3	1.42 (0.48, 3.85)	1.02 (0.27, 3.33)	1.09 (0.24, 4.40)
School 4	0.38 (0.12, 1.05)	0.78 (0.20, 2.76)	0.47 (0.10, 2.05)
School 5	1.07 (0.29, 3.56)	0.79 (0.15, 3.49)	0.80 (0.11, 4.79)
<b>Network</b>			
$\tau_r$	0 (0, $6.6 \times 10^{-53}$ )	0 (0, $1.4 \times 10^{-60}$ )	0 (0, $2.3 \times 10^{-62}$ )
$\gamma_r$	0.419 (0.02, 0.92)	0.415 (0.02, 0.92)	0.417 (0.02, 0.92)
$\Sigma_{rr}$	6.21 (3.29, 11.01)	10.32 (5.49, 19.59)	14.92 (7.12, 27.76)

odds of partaking in each adverse health behaviour. Here the baseline level is mostly A's, and decreasing the grade category almost always exhibits an increased and significant odds ratio. For example, adolescents who score mostly C's or below have significant odds ratios of 8.05 (alcohol), 5.46 (cigarettes) and 14.66 (marijuana), when compared to the baseline mostly A's category. Finally, the table shows that after accounting for all the other components in the model, there are no statistically significant school effects for schools 2, 3, 4 and 5 when compared to the reference level school 1, with all 95% credible intervals across the three responses containing the null odds ratio of 1.

#### 4.4. Peer network effects

Table 4 provides the posterior means and 95% credible intervals for the variances relating to the network random effects in the model, which quantify the variation among the individual friendship network effects. The posterior means for alcohol, cigarettes and marijuana are respectively 6.21, 10.32 and 14.92, suggesting that the greatest level of variation is for marijuana. This finding is confirmed by the posterior probability that  $\mathbb{P}(\Sigma_{33} > \Sigma_{11} \cap \Sigma_{33} > \Sigma_{22}) = 0.892$ , suggesting a very clear size ordering among these variances with the variance relating to the marijuana response being the largest.

The isolation random effects for not nominating a friend are denoted by  $u_1^*$ ,  $u_2^*$ , and  $u_3^*$ , and on the odds ratio scale their estimates and 95% credible intervals are given by: Alcohol - 2.31 (1.34, 4.06); Cigarettes - 3.32 (1.88, 6.11); and Marijuana - 4.57 (2.39, 9.39). All these estimates and 95% credible intervals are greater than one, suggesting that being isolated from others (i.e. not nominating a friend) increases the likelihood of drinking alcohol, smoking cigarettes, and having used marijuana. The posterior mean

marijuana isolation effect is the largest of the three but not significantly so, as all three 95% credible intervals overlap. However, that said there is relatively strong evidence of a clear size ordering in these isolation effects, because the model produced the following posterior probabilities:  $\mathbb{P}(u_3^* > u_1^* \cap u_3^* > u_2^*) = 0.793$ ,  $\mathbb{P}(u_3^* > u_1^*) = 0.962$ , and  $\mathbb{P}(u_3^* > u_2^*) = 0.806$ . Thus it appears that isolation has the largest effect on the marijuana response.

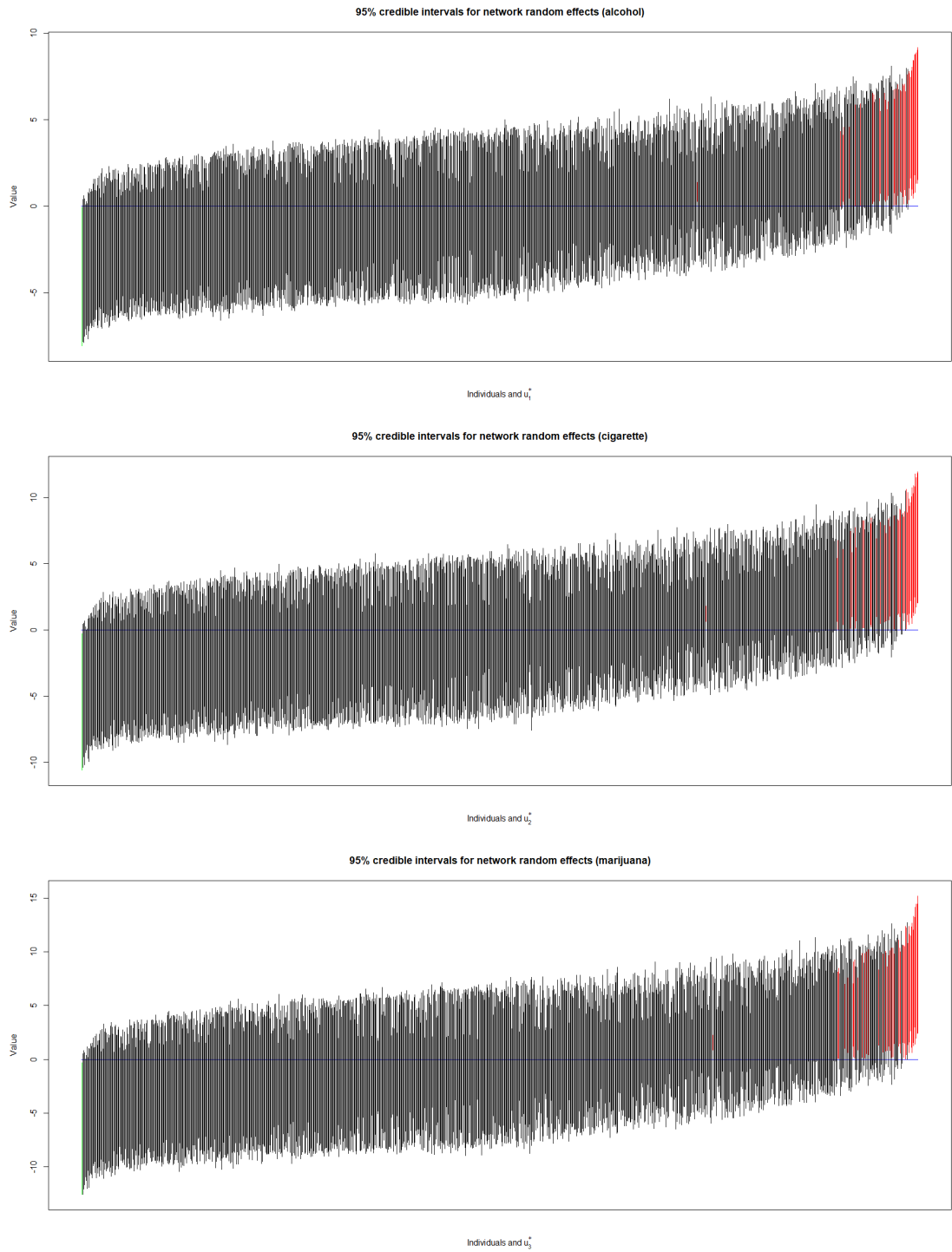
Figure 3 displays the 95% credible intervals for the individual friendship random effects and the isolation effect relating to each response, namely  $\{\mathbf{u}_1, u_1^*\}$  (alcohol),  $\{\mathbf{u}_2, u_2^*\}$  (cigarettes), and  $\{\mathbf{u}_3, u_3^*\}$  (marijuana). The effects are ordered by posterior mean on the horizontal axis, and those in black are not significantly different from zero at the 5% level. The instances in green are significantly different from zero at the 5% level and contain only negative values. There was only one case of this for each of the three responses, all attributable to the same individual. Thus, holding everything else equal, having nominated this individual as a friend was observed to have decreased the likelihood of drinking alcohol, smoking cigarettes, and having used marijuana. Those in red are significantly different from zero at the 5% level and contain only positive values. There were 38, 41, and 43 instances of this relating to the alcohol, cigarette, and marijuana response respectively, and nominating these individuals as friends increases ones propensity to smoke, drink, and use marijuana.

#### 4.5. Spatial effects

The spatial standard deviation ( $\tau_r$ ) and dependence ( $\gamma_r$ ) parameters are displayed in Table 4. The table shows that the spatial effect is essentially non-existent, as the posterior means for the conditional standard deviations  $\tau_r$  are almost zero for all three adverse health behaviours, and the upper limit of the 95% credible interval is also very close to zero. This is further emphasised by the posterior means for the sets of spatial random effects  $(\phi_1, \phi_2, \phi_3)$ , which range between  $-3.69 \times 10^{-53}$  and  $5.51 \times 10^{-53}$ . This lack of a spatial effect after covariate and friendship network effects have been accounted for confirms the overall model fit results from Table 2, which shows that the DIC values for models  $\mathcal{M}6$  and  $\mathcal{M}8$  are almost identical.

## 5. Discussion

This paper has proposed a novel spatio-network model for binary multiple health behaviour data, which jointly captures the potential effects of covariate factors, spatial location, friendship network effects and within-individual correlations between outcomes. The model is fitted in a Bayesian paradigm using MCMC simulation, and software in the form of R code is available through GitHub at <https://github.com/GNG3/modelSoftwareMVBSN> for the purpose of reproducible research. The main advantage of our model over existing alternatives is its flexibility in being able to capture this wide range of drivers of adolescent health behaviours, where as existing models only account for a subset of them. This has allowed us to examine which of these drivers are the most important for explaining an adolescent's propensity to drink alcohol or smoke cigarettes or marijuana, and in our California-based study we have obtained a number of interesting findings.



**Fig. 3.** The 95% credible intervals for the network effects relating to each response,  $\{u_1, u_1^*\}$  (alcohol - top),  $\{u_2, u_2^*\}$  (cigarettes - middle), and  $\{u_3, u_3^*\}$  (marijuana - bottom). These effects are ordered by size, and those non-significant effects with 95% credible intervals that contain 0 are shown in black, those that contain values strictly less than 0 are shown in green, and those that contain values strictly greater than 0 are shown in red.

Our main finding is that peer effects play a large role in determining whether adolescents partake in adverse health behaviours, as their addition to the null model leads to the greatest reduction in DIC when compared to just adding either the covariates or spatial component. Furthermore, once friendship effects are included in the model, there is only a small improvement in model fit when incorporating the covariates and / or spatial component. In future work it would be interesting to see whether the relative importances of the three components are mirrored in terms of their out-of-sample predictive ability, for example using a cross validation type approach. For the covariates, the only consistently significant effect on the participation in adverse health behaviours was school exam performance, with students having poorer exam results observed to be more likely to partake in these behaviours. In contrast, the spatial effect was essentially non-existent.

Our second main finding is that the effect that a friend has on an adolescent is strongly correlated across the three binary responses, with estimated pairwise correlations ranging between 0.956 and 0.975. These correlations support the notion of co-occurrence of risky behaviours in adolescents found by [Hale and Viner \(2016\)](#). Among the three pairs of jointly modelled friendship random effects, the results show that the pair relating to the cigarette and marijuana responses are the most correlated, although the correlations for the remaining pairs are only slightly smaller.

There are a number of avenues of future work that naturally result from this paper. The first of these is the development of a comprehensive R package that allows users to fit a range of univariate and multivariate spatio-network models, which will greatly increase the usability and hence influence of the general model class proposed here, allowing social scientists and other researchers to apply these models to their own data. Secondly, as the focus of the present study was binary health behaviours, the model developed was based on a binary logistic regression structure. However, spatio-network data are not necessarily binary, and we thus intend to extend the model class to include continuous and count based responses, with exam grade point averages being a natural example of a continuous response that may exhibit peer and spatial effects. A further avenue of future work is to extend the spatial model if, unlike the present study, the data suggests that space has a sizeable effect on the outcome being modelled. Such extensions could be to examine the effects of changing the specification of the neighbourhood matrix to see what impact this has on the results, as well as allowing for between outcome correlations via a multivariate CAR type model. Finally, a spatio-network interaction involving the sets of spatial and friendship network random effects could be explored, as it may be of interest to study whether friendship effects differ depending on the Zip Code in which an individual lives.

## **Acknowledgements**

The authors gratefully acknowledge the helpful comments of the editorial team and two reviewers, which have improved the content of and motivation for this work. This research was funded by a Lord Kelvin/Adam Smith (LKAS) PhD Scholarship at the University of Glasgow.



**References**

- Bandura, A. and Walters, R. H. (1977) *Social learning theory*, vol. 1. Prentice-hall Englewood Cliffs, NJ.
- Banerjee, S., Carlin, B. and Gelfand, A. (2004) *Hierarchical Modelling and Analysis for Spatial Data*. Boca Raton: Chapman & Hall.
- Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics*, **43**, 1–59.
- Bivand, R., Pebesma, E. and Gomez-Rubio, V. (2013) *Applied Spatial Data Analysis with R*. Springer, New York, 2 edn.
- Brown, L. D., Bandiera, F. C. and Harrell, M. B. (2019) Cluster Randomized Trial of Teens Against Tobacco Use: Youth Empowerment for Tobacco Control in El Paso, Texas. *American Journal of Preventive Medicine*, **57**, 592–600.
- Browne, W. J., Goldstein, H. and Rasbash, J. (2001) Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, **1**, 103–124.
- Castillo, J. M., Jivraj, S. and Ng Fat, L. (2017) The regional geography of alcohol consumption in England: Comparing drinking frequency and binge drinking. *Health & Place*, **43**, 33–40.
- Doll, R. and Hill, A. B. (1950) Smoking and Carcinoma of the Lung. *British Medical Journal*, **2**, 739–748.
- (1954) The Mortality of Doctors in Relation to Their Smoking Habits. *British Medical Journal*, **1**, 1451–1455.
- Eddelbuettel, D. and François, R. (2011) Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, **40**, 1–18. URL: <http://www.jstatsoft.org/v40/i08/>.
- Escobedo, L. G. and Peddicord, J. P. (1996) Smoking prevalence in US birth cohorts: The influence of gender and education. *American Journal of Public Health*, **86**, 231–236.
- Fujimoto, K. and Valente, T. W. (2015) Multiplex congruity: Friendship networks and perceived popularity as correlates of adolescent alcohol use. *Social Science & Medicine*, **125**, 173–181.
- Gelfand, A. and Vounatsou, P. (2003) Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, **4**, 11–15.
- Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–533.
- Gelman, A. and Rubin, D. B. (1992) Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, **7**, 457–472.

- Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*, 169–193. University Press.
- Gilman, S., Martin, L., Abrams, D., Kawachi, I., Kubzansky, L., Loucks, E., Rende, R., Rudd, R. and Buka, S. (2008) Educational attainment and cigarette smoking: a causal association? *International Journal of Epidemiology*, **37**, 615–624.
- Hale, D. R. and Viner, R. M. (2016) The correlates and course of multiple health risk behaviour in adolescence. *BMC Public Health*, **16**, 458.
- Hu, M.-C., Davies, M. and Kandel, D. B. (2006) Epidemiology and Correlates of Daily Smoking and Nicotine Dependence Among Young Adults in the United States. *American Journal of Public Health*, **96**, 299–308.
- Institute of Medicine, C. o. U. t. B. o. S. a. G. D., Institute of Medicine (U.S.), Committee on Understanding the Biology of Sex and Gender Differences, Wizemann, T. M. and Pardue, M. L. (2001) *Exploring the Biological Contributions to Human Health Does Sex Matter?* Washington, D.C.: National Academy Press.
- Jack, E., Lee, D. and Dean, N. (2019) Estimating the changing nature of Scotland’s health inequalities by using a multivariate spatiotemporal model. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, **182**, 1061–1080.
- Jin, X., Banerjee, S. and Carlin, B. (2007) Order-free co-regionalized areal data models with application to multiple-disease mapping. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **69**, 817–838.
- Lantz, P., Jacobson, P., Warner, K., Wasserman, J., Pollack, H., Berson, J. and Ahlstrom, A. (2000) Investing in youth tobacco control: A review of smoking prevention and control strategies. *Tobacco Control*, **9**, 47–63.
- Lee, D. (2011) A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, **2**, 79–89.
- Lee, D. and Lawson, A. (2016) Quantifying the spatial inequality and temporal trends in maternal smoking rates in Glasgow. *Annals of Applied Statistics*, **10**, 1427–1446.
- Leroux, B. G., Lei, X. and Breslow, N. (2000) Estimation of Disease Rates in Small Areas: A new Mixed Model for Spatial Dependence. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (eds. M. E. Halloran and D. Berry), The IMA Volumes in Mathematics and Its Applications, 179–191. New York, NY: Springer.
- Leung, R. K., Toumbourou, J. W. and Hemphill, S. A. (2014) The effect of peer influence and selection processes on adolescent alcohol use: A systematic review of longitudinal studies. *Health Psychology Review*, **8**, 426–457.
- Lewit, E., Hyland, A., Kerrebrock, N. and Cummings, K. (1997) Price, public policy and smoking in young people. *Tobacco Control*, **6**, S17–S24.
- Lorant, V. and Tranmer, M. (2019) Peer, school, and country variations in adolescents’ health behaviour: A multilevel analysis of binary response variables in six European cities. *Social Networks*, **59**, 31–40.

- MacNab, Y. (2016) Linear models of coregionalization for multivariate lattice data: a general framework for coregionalized multivariate car models. *Statistics in Medicine*, **35**, 3827–3850.
- Martinez-Beneito, M. (2013) A general modelling framework for multivariate disease mapping. *Biometrika*, **100**, 539–553.
- National Toxicology Program (2016) National Toxicology Program: 14th Report on Carcinogens. <https://ntp.niehs.nih.gov/go/roc14>.
- Pearce, J., Hiscock, R., Moon, G. and Barnett, R. (2009) The neighbourhood effects of geographical access to tobacco retailers on individual smoking behaviour. *Journal of Epidemiology & Community Health*, **63**, 69–77.
- R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: <http://www.R-project.org/>.
- Simantov, E., Schoen, C. and Klein, J. (2000) Health-compromising behaviors: Why do adolescents smoke or drink?: Identifying underlying risk and protective factors. *Archives of Pediatrics Adolescent Medicine*, **154**, 1025–1033.
- Spiegelhalter, D., Best, N., Carlin, B. and Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **64**, 583–639.
- Starkey, F., Audrey, S., Holliday, J., Moore, L. and Campbell, R. (2009) Identifying influential young people to undertake effective peer-led health promotion: The example of A Stop Smoking In Schools Trial (ASSIST). *Health Education Research*, **24**, 977–988.
- Tranmer, M., Steel, D. and Browne, W. J. (2014) Multiple-membership multiple-classification models for social network and group dependences. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **177**, 439–455.
- United States Surgeon General (2014) The Health Consequences of Smoking – 50 Years of progress: A Report of the Surgeon General: (510072014-001).
- Valente, T. W. (2012) Network Interventions. *Science*, **337**, 49–53.
- Valente, T. W., Fujimoto, K., Unger, J. B., Soto, D. W. and Meeker, D. (2013) Variations in network boundary and type: A study of adolescent peer influences. *Social Networks*, **35**, 309–316.
- Volkow, N. D., Swanson, J. M., Evins, A. E., DeLisi, L. E., Meier, M. H., Gonzalez, R., Bloomfield, M. A. P., Curran, H. V. and Baler, R. (2016) Effects of Cannabis Use on Human Behavior, Including Cognition, Motivation, and Psychosis: A Review. *JAMA Psychiatry*, **73**, 292–297.