Geospatial Health 7(2), 2013, pp. 169-182

Spatial scale effects in environmental risk-factor modelling for diseases

Ram K. Raghavan[1], Karen M. Brenner[2], John A. Harrington Jr.[3], James J. Higgins[4], Kenneth R. Harkin[2]

[1]Kansas State Veterinary Diagnostic Laboratory, College of Veterinary Medicine, Kansas State University, Manhattan, KS 66506, USA; [2]Department of Clinical Sciences, College of Veterinary Medicine, Kansas State University, Manhattan, KS 66506, USA; [3]Department of Geography, College of Arts and Sciences, Kansas State University, Manhattan, KS 66506, USA; [4]Department of Statistics, College of Arts and Sciences, Kansas State University, Manhattan, KS 66506, USA

Abstract. Studies attempting to identify environmental risk factors for diseases can be seen to extract candidate variables from remotely sensed datasets, using a single buffer-zone surrounding locations from where disease status are recorded. A retrospective case-control study using canine leptospirosis data was conducted to verify the effects of changing buffer-zones (spatial extents) on the risk factors derived. The case-control study included 94 case dogs predominantly selected based on positive polymerase chain reaction (PCR) test for leptospires in urine, and 185 control dogs based on negative PCR. Land cover features from National Land Cover Dataset (NLCD) and Kansas Gap Analysis Program (KS GAP) around geocoded addresses of cases/controls were extracted using multiple buffers at every 500 m up to 5,000 m, and multivariable logistic models were used to estimate the risk of different land cover variables to dogs. The types and statistical significance of risk factors identified changed with an increase in spatial extent in both datasets. Leptospirosis status in dogs was significantly associated with developed high-intensity areas in models that used variables extracted from spatial extents of 500-2000 m, developed medium-intensity areas beyond 2,000 m and up to 3,000 m, and evergreen forests beyond 3,500 m and up to 5,000 m in individual models in the NLCD. Significant associations were seen in urban areas in models that used variables extracted from spatial extents of 500-2,500 m and forest/woodland areas beyond 2,500 m and up to 5,000 m in individual models in Kansas gap analysis programme datasets. The use of ad hoc spatial extents can be misleading or wrong, and the determination of an appropriate spatial extent is critical when extracting environmental variables for studies. Potential work-arounds for this problem are discussed.

Keywords: spatial extent, modifiable areal unit problem, geographical information system, leptospirosis, canine.

Introduction

The use of geographical information systems (GIS) together with remote sensing and spatial analytical models is highly relevant to animal and public health research and its applications in said areas are quite broad in their scope ranging from real-time tracking/surveillance of animal diseases, risk assessment and in assessing prevention strategies for diseases. A plethora of recent studies in medical, veterinary and public health research have employed geospatial analysis to solve different problems (Durr and Gatrell, 2004; Meade and Emch, 2010), and among them are some case-control studies that aim to identify risk factors of different diseases or health conditions from the surrounding environment. One among the many analytical methods employed to achieve this is through the extraction of environmental variables that are potential risk factors of a disease from data products of remotely sensed images (land cover/land use, elevation, soil survey), which in turn are modelled using logistic or other forms of regressions to estimate the strength of their association with case status. One can frequently find such studies to have extracted environmental variables from an area within one circular buffer zone (spatial extent) surrounding geocoded locations (Cringoli et al., 2004; Gibbs et al., 2006; Gouveia and Prado, 2010; Richards et al., 2010; Raghavan et al., 2011, 2012a). Two or more spatial extents have also been used, although relatively sparingly (Ghneim et al., 2007; Charoenpanyanet and Chen, 2008; Mutuku et al., 2009).

Corresponding author:
Ram K. Raghavan
Kansas State Veterinary Diagnostic Laboratory
College of Veterinary Medicine, Kansas State University
Manhattan, KS 66506-5701, USA
Tel. +1 785 532 5618; Fax +1 785 532 3502
E-mail: rkraghavan@vet.k-state.edu

The use of a single buffer zone for conducting such studies could be problematic. It is widely recognised among ecologists that spatial patterns observed on a landscape, and therefore their representations in GIS datasets, are spatially dependent. In other words, many of the environmental objects represented as distinct (e.g. water bodies) or continuous features (e.g. precipitation) in a geographic dataset and their properties would vary dramatically over changes in distance and direction (He and Legendre, 1994; Jelinski and Wu, 1996; Wu, 2004). Therefore, at different spatial scales, the measures of any particular phenomena in those datasets (e.g. total area, percent cover) will differ as well. It has been long recognised that among some of the methodological problems encountered when using geospatial analysis in general, including its application in spatial epidemiology, the most common are those that are associated with spatial scales (Fotheringham and Wong, 1991; Sexton, 2008). The term "scale" has been used in multiple contexts to mean different concepts and their meanings are not interchangeable. A thorough review on the different usage of this term and their interpretations can be found discussed in Withers and Meentemeyer (1999) and Dungan et al. (2002). In an ecological context, spatial scale could indicate two properties: spatial resolution, often referring to the degree of geographic detail or granularity in a dataset considered for a study, and spatial extent, which indicates the total size of a study area (Turner et al., 1989; Turner, 1990), which will be the topic of interest in this study.

It has been known for almost a century now that when data from one spatial extent is progressively aggregated into fewer and larger extents for analysis, variation in statistical results will occur. This problem is called the modifiable areal unit problem (MAUP) (Gehlke and Biehl, 1934; Openshaw, 1984; Jelinski and Wu, 1996) and currently there is no work-around for fully avoiding its effects and a general solution may be considered elusive (McMaster and Sheppard, 2004; Paez and Scott, 2004). Many mitigation strategies for MAUP have been proposed; however many such solutions are context specific (Parenteau and Sawada, 2011) and may not be directly relevant to environmental risk factor analysis for diseases. The use of more than one biologically relevant spatial extent and sensitivity analysis of variables derived from those spatial extents has been advocated instead of attempting to correct for MAUP (Fotheringham, 1989) or choosing spatial extents for a study in an ad hoc manner.

In order to evaluate the effects of MAUP on environmental risk factor analysis for diseases, a case-control study using canine leptospirosis data from the Kansas and Nebraska region and spatial variables from two environmental datasets, the National Land Cover Dataset (NLCD) (MRLC, 2013) and Kansas gap analysis programme (GAP) (KARS, 2011), was conducted. The associations between canine leptospirosis and many environmental variables in North America have been well documented, including newly urbanised areas (Ward et al., 2004), urban areas (Raghavan et al., 2011, 2012a), and water bodies and wetland areas (Ghneim et al., 2007; Raghavan et al., 2012b). Cultivated agricultural land (Kuriakose et al., 1997), forest and wooded areas (Zhang, 1988; Nuti et al., 1993), and the act of working in flooded agricultural field and forests (Sharma et al., 2006; Kawaguchi et al., 2008) have also been shown to be significantly associated with canine or human leptospirosis status from other parts of the world. Many of these environmental factors are available either directly or in a modified form in the NLCD and Kansas GAP datasets for the study region.

The objective of this study was to test if varying spatial extents (MAUP) changed the types and statistical significance of environmental risk factors of canine leptospirosis derived from land cover/land use datasets. Common mitigation strategies for avoiding MAUP are discussed and recommendations for mitigating MAUP in the context of environmental risk factor analysis for diseases are presented.

Materials and methods

Case selection

Medical records of all dogs from Kansas and Nebraska that had urine polymerase chain reaction (PCR) testing for leptospirosis performed at the Kansas State University Veterinary Diagnostic Laboratory (KSVDL) between February 2002 and December 2009 were retrospectively reviewed. When available, additional information was included, specifically results of serology and urine culture for leptospirosis. A case was defined by a positive PCR test and/or one of the following results: isolation of leptospires on urine culture, a single reciprocal serum titer ≥12,800, or a four-fold rise in the reciprocal convalescent serum titer. Thus, any of these outcomes was considered sufficient to indicate leptospirosis in the dog. If the urine PCR was negative and the reciprocal serum titers were <400, the dog would be deemed a control animal.

Molecular diagnostic testing

Urine samples for PCR were handled for DNA isolation as previously reported (Harkin et al., 2003a). DNA samples were subjected to the semi-nested, pathogenic *Leptospira* PCR assay described by Woo et al. (1997) that amplifies a conserved region of the 23S rDNA, with minor modifications. A unique Taqman probe was incorporated to distinguish pathogenic *Leptospira* from saprophytic serovar varieties of this spirochete. This test has been commercially available through the KSVDL since 2002.

Serological testing

The microscopic agglutination test was performed on all blood samples submitted to the KSVDL for leptospiral serological testing. The test was performed for serovars Canicola, Bratislava, Pomona, Icterohemorrhagiae, Hardjo and Grippotyphosa.

Leptospiral culture

Urine culture was performed by inoculating 1 ml of urine obtained by cystocentesis immediately into 10 ml of liquid Ellinghausen-McCullough (EM) media, gently vortexing this inoculation and transferring 1 ml of this into another 10 ml of liquid EM media. One milliliter of each dilution (1:10 and 1:100) was then subsequently inoculated into separate 10 ml of semi-solid EM media. All tubes were incubated at 30 °C in an ambient atmosphere incubator and evaluated for evidence of growth weekly.

Demographic information

Medical records were reviewed to obtain the following information: the patient's age, rounded up to the nearest month, at the time of sample submission; the date of sample submission; and the client's street address at the time of sample submission.

Geocoding

For the purposes of this study, it was assumed that dogs spent most of their lives with their owner's residences since the information regarding their movement was difficult to obtain and when available such information was based on the subjective recollection of their owners, a potential source error. Household addresses with information pertaining to house number, street, city, state and zip code were provided by clients at the time specimens for leptospirosis testing were submitted. Addresses were retrospectively verified for their accuracy either by using MapQuest (Map Quest; America Online, Denver, USA) or Google Maps (Google Inc.; Mountain View, USA) and/or calling telephone numbers provided by clients. Geographic coordinates for these addresses were derived using a Geocode tool in ArcMap (version 9.3.1) software (Environmental Systems Research Institute, Redlands, USA) and US Census 2007 TIGER (Topographically Integrated Geographic Encoding and Referencing system) shapefile with street level address information (US Census Bureau, 2010). The geographic coordinates for unmatched addresses were obtained using Google Earth software (version 5.2.1.1329) (Google Inc.; Mountain View, USA). In all, geographic coordinates for 94 cases (out of 97) and 185 (out of 195) control data points in Kansas and Nebraska were obtained.

Projection and data storage

All GIS data used in this study were projected (or reprojected from their original spatial reference) in USA Contiguous Equal Area Conic Projection that is based on the Geographic Coordinate System North American 1983 Geographic Datum. The choice of projection system was influenced by the types of spatial analysis performed as it was essential to maintain accurate area measurements of land cover types surrounding case/control locations. All original, intermediate and processed GIS data were stored in a SQL Server (version 2008) (Microsoft Corporation, Redmond, USA) and ArcSDE (version 9.3.1) (Environmental Systems Research Institute, Redlands, USA).

Host factors

Observations were grouped into five age groups <1 years, 1-4 years, 4-7 years, 7-10 years and >10 years; two sexes and individual breeds were kept without grouping as a categorical variable.

Land cover variables

The publicly available 2001 National Land Cover Dataset (NLCD) (MRLC, 2013) (Homer et al., 2007; Wickham et al., 2010) for the study region was obtained from the United States Geological Survey (USGS) in a raster grid format. Land cover grids surrounding individual case/control locations were

extracted from the raster dataset using 5,000 m polygon buffers, and converted to polygon area features in ArcMap. Ten incremental circular buffers each of 500 m were then constructed around individual case/control locations that represented different spatial extents up to a maximum distance of 5,000 m. The buffers were overlaid with NLCD data one incremental buffer at a time, and the area of land cover types within each buffer was computed (Fig. 1). The area of different land cover type within individual buffers was divided by the total area of the respective buffer to generate percent land cover values. The process of quantifying cumulative land cover percentages was automated using a geoprocessing script written in the Python 2.4 scripting language in ArcMap.

Land cover percentages surrounding case/control locations at incremental distances were also derived using Kansas GAP data (KARS, 2010) with case/control locations located completely within Kansas. Land cover information surrounding case/control locations within the state of Nebraska was publicly available in the form of a GAP dataset (NE GAP, 2010); however, a separate analysis with Nebraska data was not conducted due to concerns of potential over-fitting of logistic models with fewer cases (n = 27) and controls (n = 29) in relation to the total number of land cover variables (n = 16).

Statistical analysis

All statistical procedures were performed using the R statistical package (R Core Development Group, 2011), and all numerical data were originally stored and organised for statistical analysis in Microsoft Excel 2010 (Microsoft Corporation, Redmond, USA).

The effect of host factors including age group (<1 year as reference category), sex (female as reference category) and breed (unknown breed as reference category) were analysed individually by fitting univariable logistic regressions.

Odds ratios (ORs) and their 95% confidence intervals (CIs) derived using logistic regressions were used to determine the risks associated with land cover variables to dogs. Land cover variables extracted from NLCD and KS GAP datasets were grouped separately (Table 1) and analysed independently in two separate steps. Observations of all land cover variables were kept in their original measurement units (percentages) in a continuous format. Land cover variables within 500 m incremental distances were screened for their association with leptospirosis by fitting univariable logistic regressions and care was taken not to eliminate variables deemed to be clinically important (Hosmer and Lemeshow, 2000), and variables with a significance level of P <0.1 were selected for further

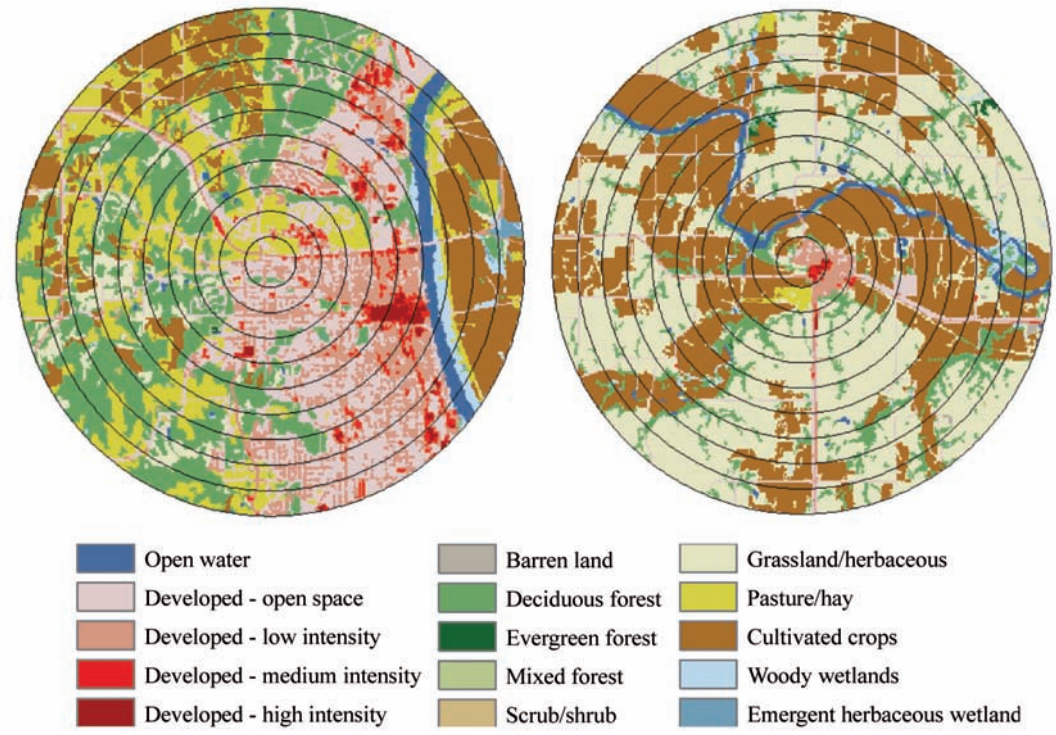

Fig. 1. Land cover/land use pattern (NLCD) surrounding a single case (left) and control (right) locations in the study region. Concentric circles are spaced 500 m apart up to 5,000 m surrounding these locations.

Table 1. Land cover types found in NLCD and Kansas GAP datasets.

Land cover/land use dataset	Land cover/land use types[d]
NLCD (source: MRLC (2010) Years: 1992-2001[a] Spatial resolution: 30 m[b] Spatial scale: 1:100,000[c]	Open water; developed - open space; developed - low intensity; developed - medium intensity; developed - high intensity; barren land; deciduous forest; evergreen forest; mixed forest; scrub/shrub; grassland/herbaceous; pasture/hay; cultivated crops; woody wetlands; and emergent herbaceous wetland.
Kansas GAP (source: KARS (2010) Years: 1995-2000 Spatial resolution: 15 m[b] Spatial scale: 1:100,000[c]	Forest/woodland (maple - basswood forest; oak - hickory forest; post oak - blackjack oak forest; pecan floodplain forest; ash - elm - hackberry floodplain forest; cottonwood; floodplain forest; mixed oak floodplain forest; evergreen forest; disturbed land; bur oak floodplain woodland; mixed oak ravine woodland; post oak - blackjack oak woodland; cottonwood floodplain woodland; deciduous woodland); shrubland (sandsage shrubland, willow shrubland, salt cedar or tamarisk shrubland); prairie (tallgrass prairie, sand prairie, western wheatgrass prairie, mixed prairie, alkali sacaton prairie, shortgrass prairie, salt marsh/prairie, low or wet prairie); marsh (freshwater marsh, bulrush marsh, cattail marsh, weedy marsh); conservation reserve programme; cultivated land; water; and urban areas.

[a]Time period during which satellite images of land cover were captured for creating the data set, including multiple images within a year
[b]The fineness of ground data as captured by a satellite image, shorter resolution meaning higher clarity
[c]The scale for which interpretations are appropriate
[d]Items within parentheses were grouped to represent broader land cover types whose names are in italics.

analysis. A multicollinearity test was conducted among screened variables by estimating the variance inflation factor (VIF) (variables with a VIF >10 considered to indicate multicollinearity) (Dohoo et al., 2003). Multivariable stepwise logistic regression models (both directions) were fitted using a significance level, P = 0.05 for variable entry and P >0.1 for a variable to be removed from the model. All models were ranked using Akaike information criterion (AIC) value and the model with lowest AIC value was deemed to be the best fitting model. The model performance was measured using deviance χ^2 goodness-of-fit test (P <0.05 indicates poor fit), and the model predictive ability was measured using the area under receiver operating characteristic curve (AUC) value. Confounding effects of host factors, age group of dogs (<1-year-old as reference category), sex (female as reference category), and breed on land cover variables were estimated by including them one at a time in the final logistic model. If such inclusion changed the coefficients of land cover variables by at least 10% or more, then the adjusted ORs were recorded from those models. The univariable screening, multivariable stepwise modelling, and checks for host-factor confounding were repeated with variables within each spatial extent and a total of 10 models for NLCD and 10 models for KS GAP datasets were derived. A Monte Carlo test based on the empirical variogram of residuals and their spatial envelopes (generated by permutations of data values across spatial locations) was used to check for residual spatial autocorrelation using the geoR library of R Statistical Package 2.11.1 (Ribeiro and Diggle, 2001; Ribeiro et al., 2003).

Results

There were 94 dogs identified as cases based on a positive PCR (n = 90 dogs), isolation of leptospires from the urine (n = 1), a single reciprocal titer ≥12,800 (n = 2), or a four-fold rise in serum reciprocal titers (n = 1) and for which geographic coordinates could be obtained. There were 185 control dogs that had a negative PCR and a reciprocal serum titer of <400 and for which geographic coordinates could be obtained. Since the time this study was conducted, all dogs diagnosed by PCR were found to be infected with serovar *Grippotyphosa* based on evaluation of variably numbered terminal repeat sequences (unpublished data). Dogs' age group, sex and breed were not significantly associated with leptospirosis status. Among 94 cases

Table 2. Urban *versus* rural geographic distribution of cases/controls in the study region.

Place*	% cases (n)	% controls (n)
Wichita	33.7 (32)	28.8 (53)
Manhattan	13.8 (13)	19.5 (36)
Lincoln	10.5 (11)	9.0 (17)
Omaha	9.5 (09)	5.2 (10)
Kansas City	6.3 (06)	4.6 (08)
Topeka	6.3 (06)	5.9 (11)
Others, rural	19.9 (17)	27.0 (50)

*Cases and controls found completely within urban boundaries of the major cities in the region were estimated in a GIS. Geographic boundary files for the cities were obtained from the US Census Bureau as a TIGER line file (US Census Bureau, 2008).

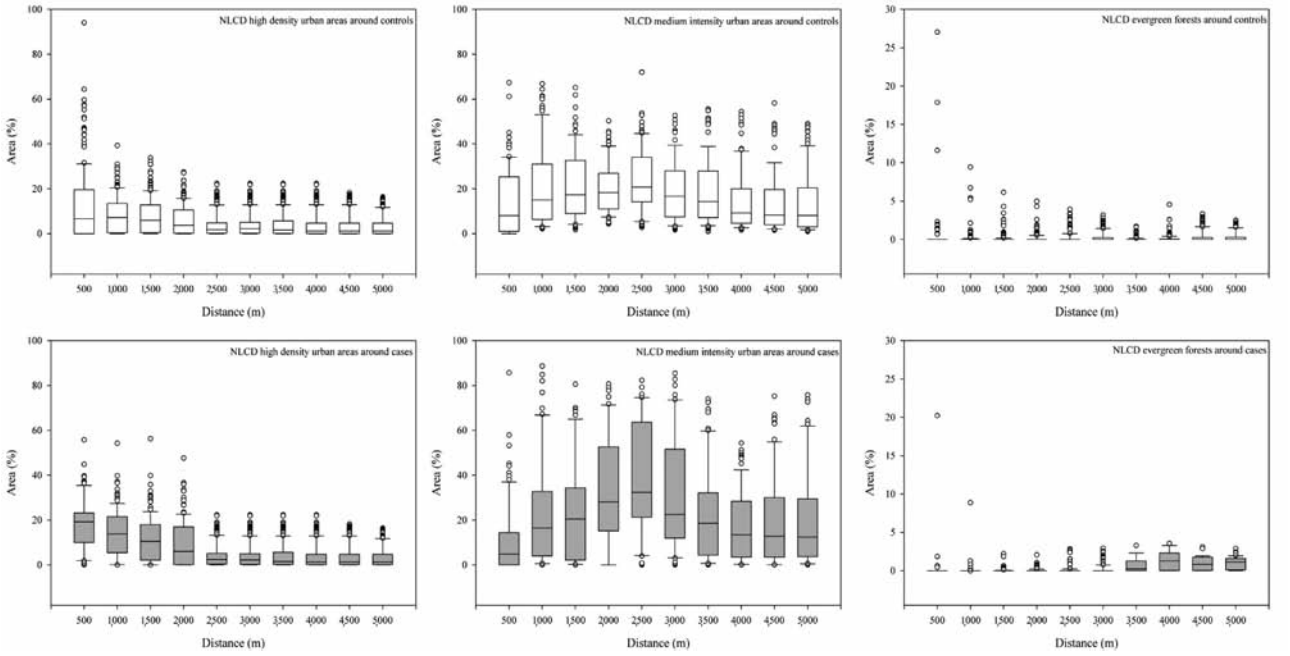

Fig. 2. Percentage area distribution of different NLCD land cover types surrounding case/control locations in the study region.

and 185 controls evaluated in this study, a majority had their physical addresses located in the major cities of the region. All remaining cases and controls had rural addresses or they were from smaller cities in the study region (Table 2).

Statistical distribution of percentage area occupied by different land cover types in NLCD and Kansas GAP datasets within incremental spatial extents are presented in Figs. 2 and 3, respectively (distributions of only those variables that were significantly associated with case status in this study are present). Two aspects of these distributions can be noticed. First, the median values of some land cover features (e.g. different urban areas) decrease with an increase in distance and other features (e.g. agriculture, forest/woodland) increase with distance. Second, a noticeable difference in the statistical distribution among variables can be seen at some spatial extents but only minimal or no

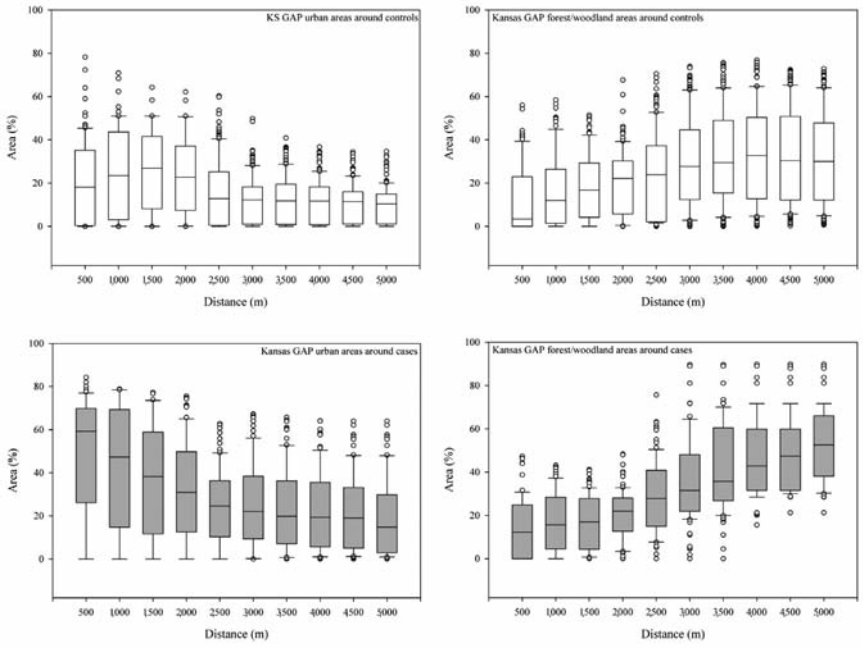

Fig. 3. Percentage area distribution of different KSGAP land cover types surrounding case/control locations in the study region.

differences at other spatial extents. For instance, the differences in distribution for high density urban area for cases and controls are readily evident at smaller distances up to 2,000 m and then the difference tapers off beyond that point. These aspects of variable distribution show an overall change in the spatial composition for land cover variables as a function of distance, in addition to their differences surrounding case/control locations.

Results of the multivariable logistic regression with NLCD land cover variables (Table 3) indicated changes to the statistical significance and types of risk factors identified as the spatial extents increased around cases/controls. Dogs were at a significantly higher risk from land cover areas represented by developed high intensity areas at all spatial extents up to 2,000 m. However, developed medium intensity was the only land cover feature statistically significant when the spatial extent reached 2,500 m and up to 3,000 m. From 3,500 to 5,000 m evergreen forests was only the significant land cover feature (Fig. 4). Similarly, the results of the multivariable logistic regression with Kansas GAP land cover variables (Table 4) revealed changes to the statistical significance and types of risk factors identified as the spatial extents increased around cases/controls. Dogs were at a significantly higher risk from land cover areas represented by urban areas for all spatial extents surrounding their homes up to 2,500 m. Forest

Table 3. Results of multivariable logistic models fit within incremental distances from dogs' residences for NLCD land cover features associated with leptospirosis status in the study region (n = 94 cases, 185 controls).

Distance (m)	Land cover feature[a]	Coefficient	P-value	OR[b]	95% CI[c]	AUC[d]
500	Developed - open space	0.822	0.078	2.28	0.54 - 9.65	0.72
	Developed - high intensity	0.400	0.027*	1.49	1.19 - 1.88	
1,000	Developed - open space	0.819	0.079	2.27	0.51 - 10.06	0.71
	Developed - high intensity	0.402	0.029*	1.49	1.19 - 1.88	
	Pasture/hay	1.503	0.090	4.50	0.80 - 25.27	
1,500	Developed - open space	0.796	0.077	2.22	0.62 - 7.92	0.77
	Developed - high intensity	0.401	0.024*	1.49	1.19 - 1.88	
	Pasture/hay	1.468	0.090	4.34	0.82 - 23.06	
2,000	Developed - open space	0.790	0.078	2.20	0.64 - 7.63	0.78
	Developed - high intensity	0.404	0.027*	1.50	1.20 - 1.88	
	Developed - medium intensity	0.659	0.070	1.93	0.97 - 3.85	
	Pasture/hay	1.432	0.091	4.19	0.73 - 24.05	
2,500	Developed - high intensity	0.409	0.611	1.51	0.93 - 2.43	0.78
	Developed - medium intensity	0.624	0.016*	1.87	1.44 - 2.41	
	Pasture/hay	1.433	0.091	4.19	0.73 - 24.03	
3,000	Developed - medium intensity	0.626	0.014*	1.87	1.45 - 2.42	0.77
	Pasture/hay	1.430	0.095	4.18	0.73 - 23.77	
	Evergreen forest	0.455	0.082	1.58	0.96 - 2.59	
3,500	Developed - medium intensity	0.593	0.071	1.81	0.97 - 3.39	0.67
	Evergreen forest	0.498	0.024*	1.65	1.33 - 2.03	
4,000	Developed - medium intensity	0.588	0.075	1.80	0.98 - 3.31	0.69
	Evergreen forest	0.526	0.022*	1.69	1.37 - 2.10	
4,500	Developed - medium intensity	0.586	0.077	1.80	0.90 - 3.57	0.70
	Evergreen forest	0.527	0.021*	1.69	1.14 - 2.51	
5,000	Developed - medium intensity	0.588	0.077	1.80	0.91 - 3.58	0.71
	Evergreen forest	0.533	0.020*	1.70	1.15 - 2.53	

[a]Continuous format, presented as percentage areas within incremental distances from dogs' residences. Host factors (age, sex, breed) were kept as categorical variables when final multivariable models in each spatial extent were tested for confounding (none found)
[b]Odds ratio
[c]Low and high limits of the 95% confidence interval
[d]Area under the receiver operating characteristic curve
*Significantly associated ($p < 0.05$) with leptospirosis status.

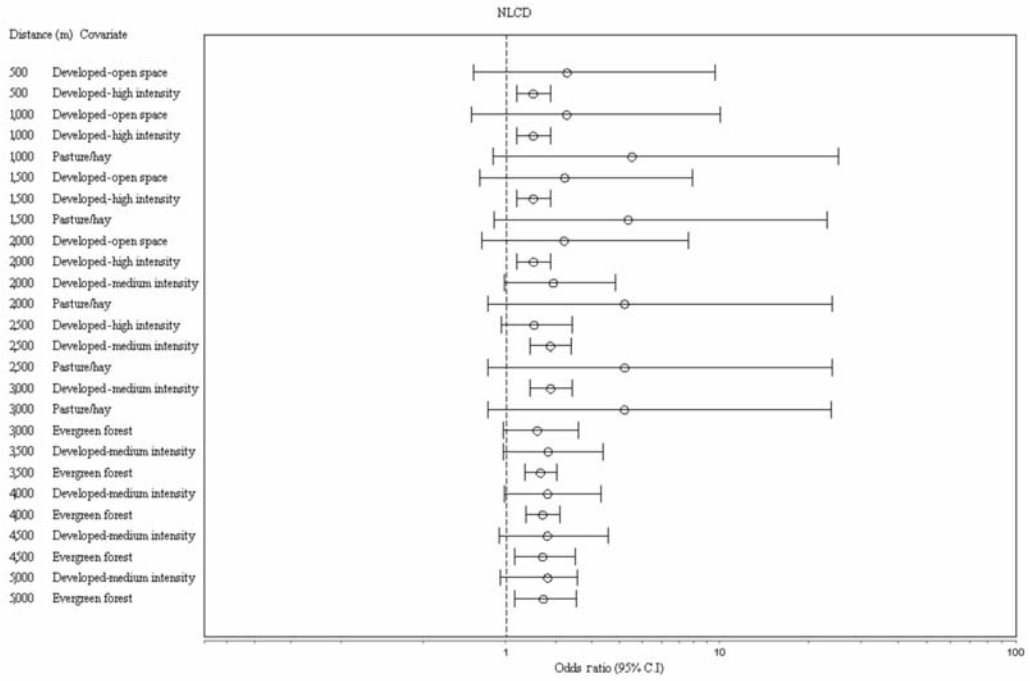

Fig. 4. Forest plot of odds ratios and 95% confidence interval (CI) of NLCD covariates retained in final multivariable logistic regression model (x-axis is in log scale).

and woodland areas was the only significant land cover feature when the spatial extent reached 3,000 m and up to 5,000 m surrounding their homes (Fig. 5). No other NLCD or Kansas GAP land cover variables were found to significantly improve the model fit when added to individual models. Host factor effects of age, gender and breed did not change the estimates of land cover variables more than 10%. The deviance goodness-of-fit test did not indicate serious model inadequacies at incremental spatial extents, and non-linearity in logit and residual autocorrelation were not noted for any models.

Model performance measured by the AUC value (for all NLCD and KS GAP models derived at incremental spatial extents) did not reveal serious flaws in model predictive ability. In terms of AUC, models in both NLCD and KSGAP categories performed moderately better when variables from spatial extents in the range of 2,000-3,000 m were used. However, the difference between the weakest performing model (0.67 at 3,500 m) and the strongest performing model (0.78 at 2,000 and 2,500 m) for NLCD and, the difference between the weakest performing model (0.72 at 5,000 m) and strongest performing model (0.83 at 2,500 m) for KS GAP variables were identical and was only a moderate 0.11. No obvious trend in AUC values with changing spatial extents could be observed (Fig. 6), although the estimates of significant variables retained in those models differed.

Discussion

The primary objective of this study was to detect if risk factors derived using candidate variables from ad hoc spatial extents changed due to the sizes of those spatial extents, and the results have confirmed such an effect. The relationship between land cover/land use and geographic distance is innate, and the noted changes in this study are potentially a reflection of the changes to the proportions and overall composition of land cover areas found within those spatial extents. Not only newer land cover types appear when spatial extents increase, also properties such as patch fragmentation, complexity, land cover/land use size and shape can greatly vary with changing spatial extents (Turner et al., 1989; Wu et al., 1997). This study has revealed that ad hoc spatial extent(s) could undermine the robustness and reliability of estimated model metrics, and limit the comparability of risk factors derived from different spatial extents. Furthermore, ad hoc spatial extent(s) could potentially result in finding biased associations, leading to incorrect disease risk factors and possibly the failure to detect important risk factors as well. While many studies have used synthetic data to demonstrate MAUP effects under different contexts (e.g. Amrhein, 1991; Saura-Martinez, 2001; Swift et al., 2008), the unique contribution of this study is the demonstration of spatial extent induced MAUP on environmental risk factors associ-

Table 4. Results of multivariate logistic models fit within incremental distances from dogs' residences for Kansas GAP (Gap Analysis Program) land cover features associated with leptospirosis status in the study region (n = 68 cases, 156 controls).

Distance (m)	Land cover feature[a]	Coefficient	P-value	OR[b]	95% CI[c]	AUC[d]
500	Urban areas	0.711	0.015*	2.04	1.37 - 3.02	0.76
	Cultivated land	1.141	0.081	3.13	0.92 - 10.59	
1,000	Urban areas	0.715	0.017*	2.04	1.38 - 3.03	0.78
	Prairie	1.811	0.090	6.12	0.89 - 42.00	
1,500	Urban areas	0.723	0.012*	2.06	1.39 - 3.06	0.77
	Prairie	1.832	0.091	6.25	0.91 - 43.06	
2,000	Urban areas	0.721	0.018*	2.06	1.38 - 3.06	0.80
	Prairie	1.835	0.091	6.27	0.89 - 44.22	
2,500	Urban areas	0.700	0.026*	2.01	1.36 - 2.99	0.83
	Prairie	1.841	0.092	6.30	0.89 - 44.48	
	Shrubland	0.892	0.068	2.44	0.89 - 6.66	
3,000	Forest/woodland	0.628	0.006*	1.87	1.45 - 2.42	0.75
	Prairie	1.841	0.094	6.30	0.73 - 54.65	
	Shrubland	0.911	0.068	2.49	0.91 - 6.78	
3,500	Forest/woodland	0.698	0.006*	2.01	1.55 - 2.6.3	0.74
	Shrubland	0.915	0.071	2.50	0.84 - 7.40	
4,000	Forest/woodland	0.698	0.004*	2.01	1.56 - 2.60	0.76
	Shrubland	0.917	0.070	2.50	0.83 - 7.50	
	Marsh	1.008	0.092	2.74	0.86 - 8.69	
4,500	Forest/woodland	0.700	0.000*	2.01	1.56 - 2.60	0.73
	Shrubland	0.918	0.074	2.50	0.83 - 7.52	
5,000	Forest/woodland	0.700	0.001*	2.10	1.55 - 2.61	0.72
	Shrubland	0.918	0.077	2.50	0.83 - 7.56	

[a]Continuous format, presented as percentage areas within incremental distances from dogs' residences. Host factors (age, sex, breed) were kept as categorical variables when final multivariable models in each spatial extent were tested for confounding (none found)
[b]Odds ratio
[c]Low and high limits of the 95% confidence interval
[d]Area under the receiver operating characteristic curve
*Significantly associated ($p < 0.05$) with leptospirosis status.

ated with disease diagnostic data received at a diagnostic facility.

The choices of ad hoc spatial extents that researchers commonly choose could be influenced by several factors, ready access to spatial analysis tools - such as the buffer analysis tool, relative convenience that such buffer features provide in terms of geospatial analysis and presentation, and also possibly a general lack of awareness of MAUP when conducting spatial analysis. Justifications for such choices are seldom provided in the literature. It is indeed difficult to determine a spatial extent that is agreeable for everyone but on the other hand, many landscape ecological studies have repeatedly shown that spatial extents do determine the range of patterns and processes that can be detected on a landscape, and have cautioned researchers of the uncertainties associated with spatial extent changes (Turner et al., 1989; Fotheringham et al., 1991; Wu, 2004). Space and therefore land cover/land use are continuous phenomena and drawing any discrete boundaries over them to extract meaningful data will introduce complications. This can be seen when studying a disease such as canine leptospirosis (recorded predominantly in urban settings) that at shorter distances the physical environment surrounding case-control locations will be dominated by variables representing the built environment, while others, such as forest and woodland areas may be seldom found and yet quite relevant (Fig. 1). A change in study area also alters the biological perspectives of a researcher since such changes are typically accompanied by changes to habitat area, habitat quality in terms heterogeneity and fragmentation. Therefore, choosing an appropriate spatial extent

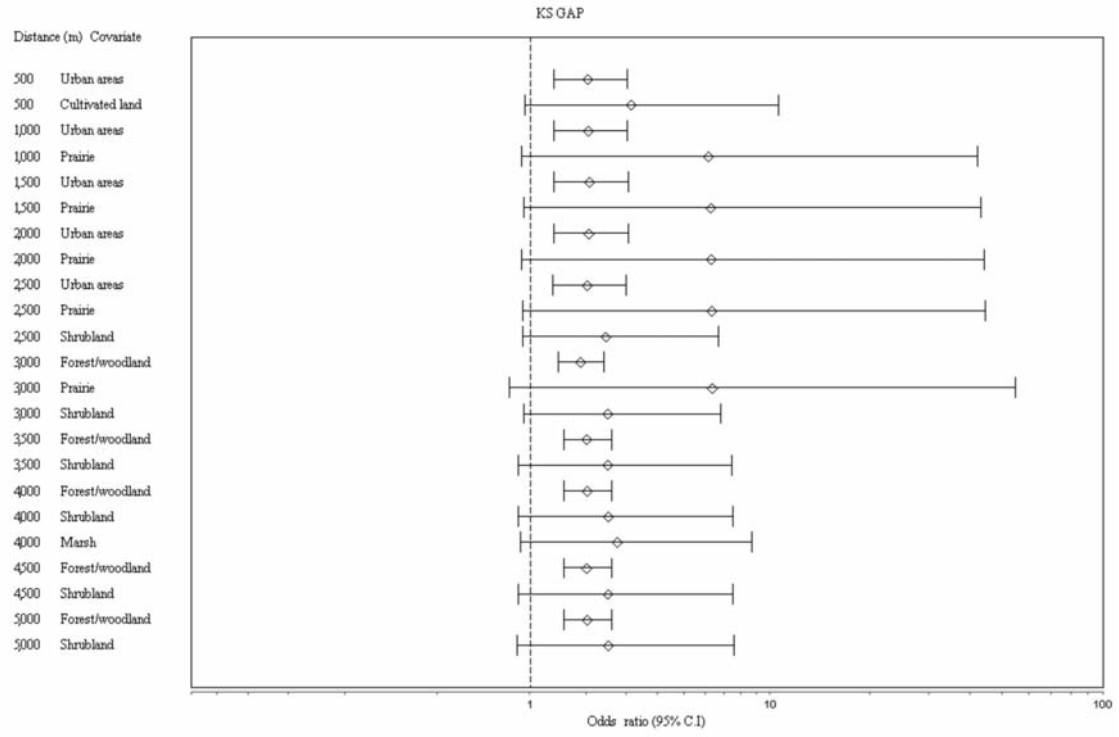

Fig. 5. Forest plot of odds ratios and 95% confidence interval (CI) of KS GAP covariates retained in final multivariable logistic regression model (x-axis is in log scale).

must be considered as a critical component of a study design.

MAUP is usually encountered when employing data that are extracted from different resolutions of spatial data (often broadly referred to as scale effect), and also due to differences in how spatial partitioning of a study region is made (zonal effect). Scale effects are expressed not only due to changes in resolution but also the extent of spatial data used in a study (Turner et al., 1989; Wu et al., 1997). Unlike most of the earlier demonstrations of MAUP by other researchers, the findings shown in this study are a result of scale effect on MAUP, particularly due to the spatial extent component of scale, not resolution. As the spatial extent gradually increases, the percent land cover values (represented by point locations of disease status) are progressively aggregated, affecting the scale of the analysis. Comprehensive discussions on different manifestations of MAUP can be found in Openshaw and Taylor (1979); Fotheringham and Wong (1991). Most discussions in the epidemiology literature regarding MAUP can be found within the realm of health data associations with socio-economic variables, often obtained from census data that are aggregated over arbitrary areal units such as zip codes or counties (zonal effects). In such circumstances, researchers have no control over how variable aggregations are made and/or how those areal units are determined. However, when MAUP is encountered due to spatial extent effects, then it may be possible to make biologically meaningful choices for the size of study area during the experimental design phase of a study.

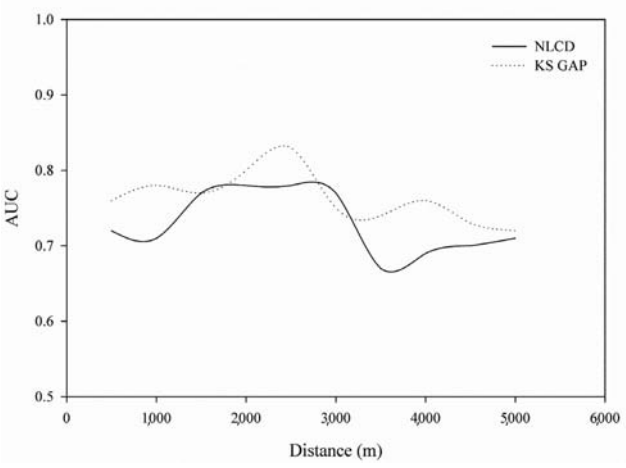

Fig. 6. Model predictive ability measured by AUC values and their trends as spatial extents increased surrounding case/control locations.

Given that MAUP is difficult to overcome and a problem that appears to be here to stay, considerable research has focused on finding ways to mitigate its

effects; although, many of the recommended approaches/work-arounds are very context specific and have mostly concentrated on zonal effects (Amhrein, 1995; Wu et al., 1997; Swift et al., 2008). The simplest way out of MAUP is for researchers to use individual level data (often referred to as basic units in the literature) since MAUP is a product of aggregation and scale dependence (Openshaw, 1984; Fotheringham, 1989). However, this may not be feasible for most situations considering the unlikely availability of such data due to privacy concerns, and is certainly not relevant to overcome the situation being discussed in this study. Openshaw (1984) in one of his four solutions to MAUP stated that MAUP will not be a predicament if researchers agreed to "objects of geographical enquiry". In other words, the focus may be placed on identifying an appropriate spatial scale that may limit the impacts of MAUP. However, it may not always be possible to identify an appropriate scale (in this case spatial extent) that would capture all the drivers of a disease such as canine leptospirosis. This can be seen from disparate studies that have shown a range of different environmental factors that are normally separated by some distances like urban areas and forest/woodland areas to be risk factors for canine leptospirosis (Ward et al., 2004; Raghavan et al., 2011; 2012a).

The use of an "optimal" spatial extent has been advocated by many in the past (Moellering and Tobler, 1972; Openshaw, 1977, 1978a, 1978b, 1984) where the goal is to artificially create a geographical structure with spatial units that has high inter-zonal variation and low intra-zonal variation. This approach has later been expounded by others using automated zoning procedures (Martin, 2001; Cockings and Martin, 2005; Haynes et al., 2007; Parenteau and Sawada, 2011). The problem with this approach however is that the determination of an optimal spatial extent even if they are computationally derived for studying disease systems could be very challenging and incomplete, considering the potential for multiple transmission pathways operating at different spatial scales and complex host and pet owner movement behaviour. Besides being limited due to its subjective nature (no one "optimal" spatial extent may be agreed upon by everyone), this method is not very practical since it is hard to define an optimal spatial extent for all the variables involved in a study (Fotheringham, 1989).

A different work-around, originally proposed by Openshaw and Taylor (1981) and later developed by Fotheringham (1989) appears to be particularly suitable for situations similar to that demonstrated in this study. Here, it is suggested not to attempt correcting MAUP itself but to acknowledge its presence in a study and instead report the sensitivity of variable relationships due to scale changes. Presumably, more faith can be placed on variables (and thereafter risk factors) that are more stable than others; and in addition this would allow researchers to first understand the magnitude of MAUP in their results, and to draw appropriate generalizations from model results prior to communications with policy makers. Openshaw (1984) recommends picking several progressively larger spatial extents for a phenomena being studied. Jelinski and Wu (1996) reasoned that, in general, observed spatial patterns in natural systems may result from factors that exert influence from multiple scales with some more obvious at some scales and others at different scales. Therefore, hierarchy theory may be useful in creating a framework for selecting spatial extents (O'Neil and King, 1998; Svancara et al., 2002; Farnsworth et al., 2006). Depending upon the cause of MAUP (zonal *versus* scale) encountered in a study, different quantification methods have been suggested in the literature for conducting sensitivity analysis and methods to identify trends that may be present due to MAUP (Knudsen and Fotheringham, 1986; Fotheringham and Wong, 1991; Jelinski and Wu, 1996). If trends are present, then conclusions can be drawn based on stable variables, and when they are absent different spatial extents may be chosen or other design considerations could be made (Haynes et al., 2007).

Appropriate spatial extents to include in a study may be explored visually using a GIS platform as a first step. Such geo-visualization could help in detecting spatial variations among candidate variables and their relationships with respect to scale changes, and to determine the spans of spatial extents (Nelson, 2001). Modern GIS software programmes are capable of providing exploratory spatial data analysis capabilities including uni/multivariate analytical methods that could also be useful in this process (Anselin et al., 2006; Parenteau and Sawada, 2011). GeogDetector (Wang et al., 2010; Wang and Hu, 2012), a software programme that uses spatial variance analysis to compare spatial consistency of disease distribution *versus* the geographical strata of environmental variables is a promising tool that could be applied to overcome MAUP effects due to changing spatial extents as well.

Conclusions

Statistical significance of disease risk factors derived using variables from spatial datasets are subject to changes to spatial extent, a problem referred to as

MAUP, and there may be policy implications if recommendations are made based on spatial epidemiological studies that ignore MAUP. A single spatial extent may not be adequate to capture scale-dependent risk factors of disease mechanisms. Potential work-around for MAUP include, but are not limited to the selection of spatial extents based on a visual analysis of spatial data in a GIS programme, and consideration of host/vector behaviour and their movement patterns, followed by sensitivity analysis of variable relations across multiple scales. We recommend studies employing spatial analysis methods to report results from all spatial extents to allow cross-comparisons with other studies.

Acknowledgements

Support for this research was partly provided by Kansas State Veterinary Diagnostic Laboratory and NSF (grant no. 0919466, Collaborative Research: EPSCoR RII Track 2 Oklahoma and Kansas: A cyber Commons for Ecological Forecasting). We express our sincere gratitude to Dr. Stewart Fotheringham, Department of Geography and Sustainable Development, University of St Andrews, Dr. Michael Sawada, Department of Geography, University of Ottawa, Dr. Michael Ward, Faculty of Veterinary Science, University of Sydney and Dr. Jingle Wu, School of Life Sciences & Global Institute of Sustainability, Arizona State University for their generous comments on earlier versions of the manuscript.

References

Amrhein C, 1995. Searching for the elusive aggregation effect: evidence from statistical simulations. Environ Plann A 27, 259-274.

Anselin L, Ibnu S, Youngihn K, 2006. GeoDa: an introduction to spatial data analysis. Geogr Anal 38, 5-22.

Charoenpanyanet A, Chen X, 2008. Satellite-based modeling of *Anopheles* mosquito densities on heterogeneous land cover in Western Thailand. The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences 27, 159-164.

Cockings S, Martin D, 2005. Zone design for environment and health studies using pre-aggregated data. Soc Sci Med 60, 2729-2742.

Cringoli G, Taddei R, Rinaldi L, Veneziano V, Musella V, Cascone C, Sibilio G, Malone JB, 2004. Use of remote sensing and geographical information systems to identify environmental features that influence the distribution of paramphistomosis in sheep from the southern Italian Apennines. Vet Parasitol 122, 15-26.

Dohoo I, Martin W, Stryhn H, 2003. Veterinary epidemiologic research. Charlottetown: AVC Inc.

Dungan JL, Perry JN, Dale MRT, Legendre P, Citron-Pousty S, Fortin MJ, Jakomulska MM, Rosenberg MS, 2002. A balanced view of scale in spatial statistical analysis. Ecography, 25, 626-640.

Durr PA, Gatrell AC, 2004. GIS and spatial analysis in veterinary science. Cambridge: CABI Publishing.

Farnsworth ML, Hoeting JA, Hobbs NT, Miller MW, 2006. Linking chronic wasting disease to mule deer movement scales: a hierarchical Bayesian approach. Ecol Appl 16, 1026-1036.

Fotheringham AS, 1989. Scale independent spatial analysis. In: The accuracy of spatial databases. 1994. Goodchild MF, Gopal S, (eds). Taylor and Francis, London, 221-228 pp.

Fotheringham AS, Wong DWS, 1991. The modifiable areal unit problem in multivariate statistical analysis. Environ Plann A 23, 1025-1044.

Gehlke C, Biehl K, 1934. Effects of grouping upon the size of the correlation coefficient in census tract material. J Am Stat Assoc 29, 169-170.

Ghneim GS, Viers JH, Chomel BB, Kass PH, Descollonges DA, Johnson ML, 2007. Use of a case-control study and geographic information systems to determine environmental and demographic risk factors for canine leptospirosis. Vet Res 38, 37-50.

Gibbs SEJ, Wimberly MC, Madden M, Masour J, Yabsley MJ, Stallknecht DE, 2006. Factors affecting the geographic distribution of West Nile virus in Georgia, USA: 2002-2004. Vector-Borne Zoonot 6, 73-82.

Gouveia N, Prado RR, 2010. Health risks in areas close to solid waste landfill sites. Rev Saude Publica 44, 859-866.

Harkin KR, Roshto YM, Sullivan JT, Purvis TJ, Chengappa MM, 2003. Comparison of polymerase chain reaction assay, bacteriologic culture, and serologic testing in assessment of prevalence of urinary shedding of leptospires in dogs. J Am Vet Med A 222, 1230-1233.

Haynes R, Daras K, Reading R, Jones A, 2007. Modifiable neighborhood units, zone design and residents' perceptions. Health Place 13, 812-825.

He F, Legendre P, 1994. Diversity pattern and spatial scale: a study of tropical rain forest of Malaysia. Environ Ecol Stat 1, 265-286.

Homer C, Dewitz J, Fry J, Coan M, Hossain N, Larson C, Herold N, McKerrow A, VanDriel JN, Wickham J, 2007. Completion of the 2001 National Land Cover Database for the conterminous United States. Photogramm Eng Rem S 73, 337-341.

Hosmer DW, Lemeshow S, 2000. Model-building strategies and methods for logistic regression. In: Applied logistic regression. Hosmer DW, Lemeshow S (eds). 2nd ed. John Wiley & Sons, New York, 91-142 pp.

Jelinski DE, Wu J, 1996. The modifiable areal unit problem and

implications for landscape ecology. Landscape Ecol 11, 129-140.

KARS, 2010. Kansas Applied Remote Sensing. Available at: http://www.kars.ku.edu/ (accessed on May 2013)

Kawaguchi L, Sengkeopraseuth B, Tsuyuoka R, Koizumi N, Akashi H, Vongphrachanh P, Watanabe H, Aoyama A, 2008. Seroprevalence of leptospirosis and risk factor analysis in flood-prone rural areas in Lao PDR. Am J Trop Med Hyg 78, 957-961.

Knudsen CD, Fotheringham AS, 1986. Matrix comparison, goodness-of-fit and spatial interaction modeling. Int Regional Sci Rev 10, 127-147.

Kuriakose M, Eapen CK, Paul R, 1997. Leptospirosis in Kolenchery, Kerala, India: epidemiology, prevalent local serogroups and serovars and a new serovar. Eur J Epidemiol 13, 691-697.

Martin D, 2001. Developing the automated zoning procedure to reconcile incompatible zoning systems. Int J Geogr Inf Sci 17, 181-196.

McMaster R, Sheppard E, 2004. Introduction: scale and geographic inquiry. In: Scale and geographic inquiry: nature. McMaster R, Sheppard E (eds). Society and Method, Blackwell.

Meade MS, Emch M, 2010. Medical geography, Third Edition, The Guilford Press, New York.

Moellering H, Tobler W, 1972. Geographical variances. Geogr Anal 4, 34-50.

MRLC, 2013. Multi-resolution land characteristics consortium. U.S. Department of the Interior, U.S. Geological Survey. Available at: http://www.mrlc.gov/index.php (accessed on May 2013).

Mutuku FM, Bayoh MN, Hightower AW, Vulule JM, Gimnig JE, Mueke JM, Amimo FA, Walker ED, 2009. A supervised land cover classification of a western Kenya lowland endemic for human malaria: associations of land cover with larval *Anopheles* habitats. Int J Health Geog 8.

NE GAP, 2010. Available at: http://www.calmit.unl.edu/gap/ (accessed on May 2013).

Nelson A, 2001. Analysing data across geographic scales in Honduras: detecting levels of organization within systems. Agr Ecosyst Environ 85, 107-131.

Nuti M, Amaddeo D, Crovatto M, Ghionni A, Polato D, Lillini E, Pitzus E, Santini GF, 1993. Infections in an alpine environment: antibodies to hantaviruses, *Leptospira*, rickettsiae, and *Borrelia burgdorferi* in defined Italian populations. Am J Trop Med Hyg 48, 20-25.

O'Neil RV, King AW, 1998. Homage to St. Michael; or, why are there so many books on scale? In: Ecological scale: theory and applications. Peterson DL, Parker VT (eds). Columbia University Press, New York.

Openshaw S, 1977. Optimal zoning systems for spatial interaction models. Environ Plan A 9, 169-184.

Openshaw S, 1978a. An empirical study of some zone design criteria. Environ Plann A 10, 781-794.

Openshaw S, 1978b. An optimal zoning approach to the study of spatially aggregated data. In: Spatial representation and spatial interaction. Masser I, Brown PJB (eds). Martinus Nijhoff: Leiden.

Openshaw S, 1984. The modifiable areal unit problem. In: Concepts and techniques in modern geography No. 38, Geo Books, Norwich. Available at: http://qmrg.org.uk/files/2008/11/38-maup-openshaw.pdf. (accessed on May 2013).

Openshaw S, Taylor P, 1979. A million or so correlated coefficients. In: Three experiments on the modifiable areal unit problem. Wrigley N, Bennet R (eds). Statistical applications in the spatial sciences. Pion, London.

Paez A, Scott D, 2004. Spatial statistics for urban analysis: a review of techniques with examples. GeoJournal 61, 53-67.

Parenteau MP, Sawada MC, 2011. The modifiable areal unit problem (MAUP) in the relationship between exposure to NO_2 and respiratory health. Int J Health Geogr 10, 58.

R Core Development Team, 2011. R: a language and environment for statistical computing, Reference Index Version 2.11.1. R Foundation for Statistical Computing, Vienna, Austria.

Raghavan R, Brenner K, Higgins J, Van der Merwe D, Harkin KR, 2011. Evaluations of land cover risk factors for canine leptospirosis: 94 cases (2002-2009). Prev Vet Med 101, 241-249.

Raghavan RK, Brenner KM, Higgins JJ, Shawn Hutchinson JM, Harkin KR, 2012a. Hydrologic risk factors of canine leptospirosis: 94 cases (2002-2009). Prev Vet Med 107, 105-109.

Raghavan RK, Brenner KM, Higgins JJ, Shawn Hutchinson JM, Harkin KR, 2012b. Neighborhood level socio-economic and demographic risk factors of canine leptospirosis: 94 cases (2002-2009). Prev Vet Med 106, 324-331.

Ribeiro PJ, Christensen OF, Diggle PJ, 2003. geoR and geoRglm: Software for model based geostatistics. In: 3rd International workshop on distributed statistical computing (DSC 2003). Hornik K, Leisch F, Zeileis A (eds). Vienna.

Ribeiro PJ, Diggle PJ, 2001. geoR: a package for geostatistical analysis. R-News. Vienna, pp. 15-18.

Richards EE, Masuoka P, Major DB, Smith M, Klein TA, Kim HC, Anyamba A, Grieco J, 2010. The relationship between mosquito abundance and rice field density in the Republic of Korea. Int J Health Geogr 9, 32.

Saura-Martinez, 2000. Sensitivity of landscape pattern metrics to map spatial extent.

Sexton K, 2008. Modifiable areal unit problem (MAUP). Encyclopedia of quantitative risk analysis and assessment. John Wiley & Sons Ltd. New York.

Sharma S, Vijayachari P, Sugunan AP, Natarajaseenivasan K, Sehgal SC, 2006. Seroprevalence of leptospirosis among high-risk population of Andaman Islands, India. Am J Trop Med Hyg 74, 278-283.

Svancara LK, Garton EO, Chang KT, Scott MJ, Zager P, Gratson M, 2002. The inherent aggravation of aggregation: an

example with elk aerial survey data. J Wildlife Manage 66, 776-787.

Swift A, Liu L, Uber J, 2008. Reducing MAUP bias of correlation statistics between water quality and GI illness. Comput Environ Urban 32, 134-148.

Turner MG, O'Neill RV, Gardner RH, Milne BT, 1989. Effects of changing spatial scale on the analysis of landscape pattern. Landscape Ecol 3, 153-162.

Turner MG, 1990. Spatial and temporal analysis of landscape patterns. Landscape Ecol 4, 21-30.

Wang JF, Li XH, Christakos G, Liao YL, Zhang T, Gu X, Zheng XY, 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun region, China. Int J Geogr Inf Sci 24, 107-127.

Wang JH, Hu Y, 2012. Environmental health risk detection with GeogDetector. Environ Model Softw 33, 114-115.

Ward MP, Guptill LF, Wu CC, 2004. Evaluation of environmental risk factors for leptospirosis in dogs: 36 cases (1997-2002). J Am Vet Med A 225, 72-77.

Wickham JD, Stehman SV, Fry JA, Smith JH, Homer CG, 2010. Thematic accuracy of the NLCD 2001 land cover for the conterminous United States. Remote Sens Environ 114, 1286-1296.

Withers MA, Meentemeyer V, 1999. Concepts of scale in landscape ecology. In: Landscape ecological analysis: issues and applications. Klopatek JM and Gardner RH (eds). Springer-Verlag, New York, USA.

Woo THS, Patel BKC, Smythe LD, Symonds M, Norris MA, Dohnt MF, 1997. Identification of pathogenic *Leptospira* genospecies by continuous monitoring of fluorogenic hybridization probes during rapid-cycle PCR. J Clin Microbiol 35, 3140-3146.

Wu J, 2004. Effects of changing scale on landscape pattern analysis: scaling relations. Landscape Ecol 19, 125-138.

Wu J, Gao W, Tueller PT, 1997. Effects of changing spatial scale on the results of statistical analysis with landscape data: a case study. Geographic Information Sciences 3, 601-609.

Wu J, Shen W, Sun W, Tueller PT, 2002. Empirical patterns of the effects of changing scale on landscape metrics. Landscape Ecol 17, 761-782.

Zhang F, 1988. Distribution and dynamics of leptospiral serovars in frontier of Yunnan province, China. Chin J Epidemiol 9, 25-28.