Supplemental Information for Appendix B

As discussed in Appendix B on page 32, special considerations must be made when applying regression techniques to spatially structured data. Once global clustering statistics determine if there are patterns of clustering within the study area, local clustering methods can identify likely clusters within that area of concern. Local cluster detection methodologies and scan statistics are useful for pinpointing the most likely geographic locations of those clusters. In cases where there is a known point-source, focused tests can be considered. Regression analysis can then be used to adjust for confounding factors such as latency in cases, mobility and demographic variables (such as age and race). Below are some general steps and underlying logic of spatial regression, drawing primarily from Waller & Gotway[1] and Fotheringam & Rogerson[2].

1. **Model the relationship between covariates and outcome data, assuming no spatial relationship.** It is possible that all of the important covariates were included in the model, and that their spatial distribution sufficiently accounts for spatial structure in the outcome.

2. **Check for residual spatial variation.** Use techniques described above (e.g. tests for spatial autocorrelation, including Moran's I) to see if there is significant spatial variation in the residuals, which are the differences between the model-predicted outcome and the observed data. If the model's residuals display no spatial variation, there's no need to make any adjustments to your model! If the model residuals are autocorrelated, the included covariates did not fully account for spatial variation in outcome. There are a few interesting properties of the model that can generate this pattern in the residuals: **(1)** There are predictive factors omitted from the model, and including them as covariates would sufficiently reduce spatial structure in the residual values. This scenario is likeliest, and the selection of covariates measured at an appropriate spatial scale for the modeled relationship is critical.[1,3] **(2)** The outcome variable itself is spatially autocorrelated. This is an expected quality of propagating phenomena, e.g., infectious disease spread, but less so of something that is expected rarely and at random, like cancer. **(3)** The model's "goodness-of-fit," or its ability to describe the relationship between covariates and observations, is spatially structured. This can come from spatially structured differences in sampling design (i.e., differences in sampling techniques or mismatch between the scale at which a covariate is sampled and the scale at which it interacts with the response).[1,3]

3. **Accounting for residual spatial variation in your model.** There are many options within the family of spatial regression models, and we suggest consulting with a statistician to choose the optimal approach for your data and goals.

   a. **Spatial autoregressive models:**

      I. **Residual structure reflects spatial autocorrelation of response values:** Spatial structure in the response is partly due to local interactions at a fine scale. To reflect this, the mean outcome of nearby neighbors is added to the model as a covariate (called a "spatially lagged variable"). The definition of "nearby" in this approach is flexible, and can be used to correct for mismatches of scale between observations and their predictors[1,4] or to explore alternative concepts of neighborhoods, such as groups of workers at certain facilities regardless of residence location.[3,4]

      II. **Residual structure reflects spatial autocorrelation of error:** Spatial error structure is added to the variance-covariance of the model.[1,2] Helpful variance-covariance adjustments require a solid understanding of a model's sources of error, and cancer epidemiological researchers should proceed with caution here.[1] Cancer case data typically deal in very small values relative to the population, cases tend to be underreported, and individuals may go undiagnosed for long periods of time. The latency between exposure events and cancer symptom progression gives the exposed population ample time to move around, so cancer outcomes may never get mapped near their exposure and healthy immigrants further dilute the signal. Models with community cancer data will have a lot of "unknown unknowns" that probably limit the usefulness of variance-covariance adjustments.

National Center for Environmental Health
Agency for Toxic Substances and Disease Registry

CS335369-A November 4, 2022

b. **Spatial mixed models:** a random effect term is included to account for correlated observations. Examples include a random effect (intercept or slope) for county that would group census tracts into counties. For linear regression (a less common example when dealing with health outcomes) the semivariogram can be used to assess residual spatial variance which can then also be used to model the residual spatial covariance structure and regression inference based on generalized least squares.[5]

c. **Spatial GEE methods:** For generalized linear regression models (e.g. logistic and Poisson regression) the method of generalized estimating equations (GEE) can be an alternative for providing regression inference in the presence of residual spatial variation.[5]

d. **Geographically weighted regression (GWR):** Broadly, GWR models spatial variance in the covariate/response *relationship*. Rather than adjusting one model to account for spatial dependency in the response at all locations, GWR fits the model at each data point—the implication here is that small scale, local processes are critical to explaining patterns on a wider scale, and the distinguishing assumption of GWR is that the covariates effects on the outcome are dependent on location.[6,7]

e. **Bayesian models** infer the probability of model parameters given prior information and the data, rather than the probability of the data fitting a known "true" distribution.[1,8,9] Bayesian hierarchical models use spatially correlated random intercepts to borrow information from neighboring regions to improve precision within small areas.

As described in Appendix B (on pages 30, 31, and 33) the following table presents spatial clustering methods, conditions for which the spatial clustering method are appropriate, software in which these methods can be performed, and additional notes relative to these methods and their performance. This table is not meant to be a comprehensive review of applications or a list of recommended applications but rather provides initial guidance for available tools both free and proprietary.

Global clustering statistics

Spatial Clustering Methods (by type)	Small Area OK	Small Number OK	Data Type	Application/Software	Notes
Besag-Newell R	No	Yes	point	ClusterSeer R *SpatialEpi/besag.newall*	Detects, locates and estimates the extent and intensity of circular clusters. Does not control for multiple testing problem; requires setting of several tuning parameters.
Difference of k-function	Yes	Yes	point	R *Ecespa/k1k2*	Compares the overall spatial patterns at a range of spatial scales of a case and control dataset (e.g. cancer cases and demographically linked at-risk control population). Assumption: Stationarity, independence.
Geary's C	No	No	point, polygon	GeoDa R *Spdep/geary.test* ESRI ArcGIS SAS *proc variogran*	A measure of spatial autocorrelation (local or global). Values range from 0–2. 0=positive autocorrelation; 1= none; 2=negative. inversely related to Moran's I.
Getis-Ord G	No	No	point, polygon	GeoDa R *Spdep/globalG.test* ESRI ArcGIS	The Hot Spot Analysis tool: calculates the statistic for each feature. The z-scores and p-values indicate statistical significance of clustering. Not reliable with less than 30 features; all features should have at least 1 neighbor; no feature should have all other features as neighbors.

Continued on the next page ►

Spatial Clustering Methods (by type)	Small Area OK	Small Number OK	Data Type	Application/Software	Notes
Global Moran's I	No	No	point, polygon	GeoDa R *ape.Moran.I* ESRI ArcGIS SAS *proc variogran*	Use spatial autocorrelation to evaluate clustering. Test under a normality assumption. Tends to underestimate autocorrelation for small samples. Does not account for population heterogeneity.
Bivariate Moran's I	Unclear	No	point, polygon	GeoDa	Measures the influence of one variable has on occurrence of over another close variable. Interpret with caution—can overestimate the spatial correlation. Assumption of normality.
k-function (Ripley's)	Yes	Yes	point	ClusterSeer R *spatstat/Kest* SAS	Overall testing of spatial pattern of point data. Able to simultaneous account for multivariate (categorial) data sets. Assumptions: Stationarity, independence.
Knox test	No	No	point	R *Surveillance.Knox M.test*	Does not distinguish between shifting populations or increased disease scenarios (depending on time period). Corrected with Knox-Mantel test. Distance/scale for cluster detection must be specified up front.
Oden's Ipop[10]	Yes	Yes	polygon	GeoDa	Adapted Moran's I to consider population size. Accounts for differences in population size across areas.
Potthoff and Whittinghill test[11]	Yes	Yes	polygon	R *Dcluster/pottwhitt*	Checks the ratio of the variance to the expected number of cases. If >1, the data are determined over-dispersed relative to the Poisson distribution.
Space-time k function[12]	No	No	point	R *splancs/stkhat*	Compares the observed spatial-temporal point pattern a similar space-time pattern that does not have space-time interaction.

Local clustering statistics—Local indicators of spatial association (LISA)

Spatial Clustering Methods (by type)	Small Area OK	Small Number OK	Data Type	Application/Software	Notes
Aneslin Moran's I	Unclear	Unclear	point, polygon	ClusterSeer GeoDa R *spdep/localmoran* ESRI ArcGIS	Local spatial autocorrelation statistic that identifies local clusters or local outliers. Significance testing of local Moran statistics can be somewhat problematic. Doesn't conform to a common distribution.
Besag-Newell R	No	Yes	point	ClusterSeer R *SpatialEpi/besag.newall*	Calculates I (the local statistic) the number of regions required to reach a certain number of cases (k) and the probability that the k cases form a cluster.
Getis-Ord Gi*	Yes	No	point, polygon	GeoDa R *spdep/localG* ESRI ArcGIS	Calculates the Getis-Ord Gi* statistic for each feature. Provides significance of cluster (hot or cold) using z-scores and p-values

Continued on the next page ►

Spatial Clustering Methods (by type)	Small Area OK	Small Number OK	Data Type	Application/Software	Notes
Local Geary's C	No	No	point, polygon	GeoDa R *usdm/lisa* ESRI ArcGIS	A measure of spatial autocorrelation. Values range from 0–2. 0=positive autocorrelation; 1= none; 2=negative. inversely related to Moran's I.

Focused clustering statistics

Spatial Clustering Methods (by type)	Small Area OK	Small Number OK	Data Type	Application/Software	Notes
Lawson and Waller Focused Test[13]	Yes	No	polygon	R *spatstat/WallerLawson.test*	Evaluates disease incidence relative to exposure based on distance.
Bithell's Linear Risk Score	No	No	polygon	R *spatstat/Bithell.test*	Sensitive to excess risk near a point-source.
Diggle's Method	No	No	polygon	R *spatstat/DiggleETAL.test*	Fits a model using maximum likelihood about a focus. The model parameters have associated significance that indicates potential increased risk.
Stone Maximum Likelihood Ratio Test	No	No	polygon	R *spatstat/Stone.test*	Identifies trend (descending) in risk with distance from the point-source
Tango's Focused Test	No	No	polygon	R *spatstat/TangoF.test*	Different tests that depend on the trend with distance from the point-source (decline trend or a peak-decline trend).

Scan clustering statistics

Spatial Clustering Methods (by type)	Small Area OK	Small Number OK	Data Type	Application/Software	Notes
Bonferroni Locally Adjusted Spatial Scan Statistic (BLASS)	No	No	point, polygon	—	Similar to GLASS, but utilizes a Bonferroni distribution rather than a Gumbel distribution.
FlexScan	Yes	Unclear	polygon	R *rflexscan*	Using Monte-Carlo method to test the spatial clusters. Results include a table and a map of clusters
Gumbel Locally Adjusted Spatial Scan Statistic (GLASS)	No	No	point, polygon	R *Scanstatistics/*	Also called the extreme value type I distribution. Used to find a maximum extreme value., Models the distribution of the max (or the min) of a number of samples of multiple distributions. Data must be organized to have Max and Min in separate bins
Kulldorf space and time scan statistic	Yes	Yes	point, polygon	—	Spatial scan statistic based on likelihood ratio associated with the number of cases inside and outside a scan window.
Kulldorf spatial scan statistic[14]	Yes	Yes	point, polygon	ClusterSeer R *Dcluster.Nagarwalla.test*	See above

Continued on the next page ►

Spatial Clustering Methods (by type)	Small Area OK	Small Number OK	Data Type	Application/Software	Notes
Multivariate Bayesian Scan Statistic (MBSS)	No	No	point, polygon	R *BenjaK/*	A stream-based event surveillance network; Detects and characterizes events when provided multiple data streams.
ClustR	Yes	Yes	point	R *ClustR*	A space-time cluster analysis R package.

Regression Models

Spatial Clustering Methods (by type)	Small Area OK	Small Number OK	Data Type	Application/Software	Notes
Bayesian geoadditive model	No	No	point	R *R2BayesX/bayesx*	—
Emperical Bayes Index (EBI)	No	No	point	R *spdep/EBImoran*	—
Hierarchical Bayes Spatial Modeling (HBSM)	No	No	point	R *CARBayes*	—
Generalized additive mixed models	No	No	point, polygon	R *MASS/gamm*	—
Geographic weighted regression	No	No	point, polygon	R *spgwr/gwr*	—
Joinpoint regression model	Yes	Yes	point	R *JPSurv/joinpoint*	—
Lee/Lawson Poisson Log-Linear Model	Unclear	Unclear	polygon	R *CARBayesST/ST.cluster*	—
Local generalized additive regression models	Unclear	Unclear	point	R *gam/gam.control*	—
Ordinary Least Squares Regression	Yes	No	point	R *caTools/OLS*	—
Penalized Likelihood Estimate	No	No	point	R *pmlr*	—
Poisson regression	Yes	No	point	R *glm*	—
Spatial error model	No	No	point	R *spatialreg/Gmerrorsar*	—
Spatial lag regression models	No	No	point	R *spatialreg/lagmess*	—
Srandom effects model (linear)	No	No	point	R *FRK/sre*	—
Structured additive regression models	Yes	Yes	point	R *R2BayesX/bayesx*	—

Non-spatial elevated case analysis

Spatial Clustering Methods (by type)	Small Area OK	Small Number OK	Data Type	Application/Software	Notes
Causal SIR (new form of SIR)	Yes	Yes	point, polygon	—	SIR analysis can help determine if the number of observed cancer cases in a geographic area is higher (SIR >1) or lower (SIR <1) than expected. Requires population data and age distribution for the area of interest. Also needs a reference population that has comparable demographics
Choropleth maps (purely descriptive)	Yes	Yes	polygon	R *sp/spplot*	A map where each color represents a category (can be a range of values for continuous data)
Heat map (purely descriptive)	Yes	Yes	point, polygon	R *heatmap*	Descriptive and mostly visual

Other Methods (classification, hierarchical, non-parametric)

Spatial Clustering Methods (by type)	Small Area OK	Small Number OK	Data Type	Application/Software	Notes
Cuzick-edwards kNN test	Yes	Yes	point	R *mc30/spatclustrkNN*	Nearest neighbor test. Samples cases and controls from a common spatial distribution. The nearest neighbor to a case in a cluster is often another case.
Jacquez k nearest neighbor	Yes	Yes	point	R *Jacquez.test*	—
Kernel density estimation	Yes	Yes	point	R *kdensity/kde*	—
M statistic Test	Unclear	Unclear	point	R *rstatix/box_m*	—
Mantel test	Unclear	Unclear	two matrices	R *Mantel.test*	—
Q statistics	Yes	Yes	point	R *gamlss/Q.stats*	—
Rogerson's Test	Yes	Yes	polygon	ClusterSeer R *npst*	—
Ward's Method	Yes	Yes	point	R *stats/hclust*	—

References

1. Waller, L.A. and C.A. Gotway, Applied spatial statistics for public health data. Vol. 368. 2004: John Wiley & Sons.
2. Fotheringham, A.S. and P.A. Rogerson, eds. The SAGE handbook of spatial analysis. 2009, Sage: Thousand Oaks, CA, USA. 528.
3. Gaspard, G., D. Kim, and Y. Chun, Residual spatial autocorrelation in macroecological and biogeographical modeling: a review. Journal of Ecology and Environment, 2019. 43(1): p. 19.
4. Anselin, L., Spatial Regression, in The SAGE Handbook of Spatial Analysis, A.S. Fotheringham and P.A. Rogerson, Editors. 2009, SAGE Publications, Ltd: London. p. 255–275.
5. Schabenberger, O. and C.A. Gotway, Statistical methods for spatial data analysis. 2017: Chapman & Hall/CRC.
6. Brunsdon, C., S. Fotheringham, and M. Charlton, Geographically Weighted Regression. Journal of the Royal Statistical Society: Series D (The Statistician), 1998. 47(3): p. 431–443.
7. Fotheringham, A.S., C. Brunsdon, and M. Charlton, Geographically weighted regression: the analysis of spatially varying relationships. 2003: John Wiley & Sons.
8. Congdon, P., Bayesian modelling strategies for spatially varying regression coefficients: A multivariate perspective for multiple outcomes. Computational Statistics & Data Analysis, 2007. 51(5): p. 2586–2601.
9. Lawson, A.B., Bayesian disease mapping: hierarchical modeling in spatial epidemiology. 2018: Chapman and Hall/CRC.
10. Oden, N., Adjusting Moran's I for population density. Statistics in Medicine, 1995. 14(1): p. 17–26.
11. Goungounga, J.A., et al., Impact of socioeconomic inequalities on geographic disparities in cancer incidence: comparison of methods for spatial disease mapping. BMC medical research methodology, 2016. 16(1): p. 1–14.
12. Xu, B., et al., Spatial and spatial-temporal clustering analysis of hemorrhagic disease in white-tailed deer in the southeastern USA: 1980–2003. Preventive veterinary medicine, 2012. 106(3–4): p. 339–347.
13. Waller, L.A. and A.B. Lawson, The power of focused tests to detect disease clustering. Statistics in Medicine, 1995. 14(21–22): p. 2291–2308.
14. Kulldorff, M., A spatial scan statistic. Communications in Statistics-Theory and methods, 1997. 26(6): p. 1481–1496.