A comparison of least squares regression and geographically weighted regression modeling of West Nile virus risk based on environmental parameters

Abhishek K Kala [1] , **Chetan Tiwari** [2] , **Armin R Mikler** [3] , **Samuel F Atkinson** Corresp. [1]

[1] Advanced Environmental Research Institute and Department of Biological Sciences, University of North Texas, Denton, Texas, United States

[2] Advanced Environmental Research Institute and Department of Geography and the Environment, University of North Texas, Denton, Texas, United States

[3] Advanced Environmental Research Institute and Department of Computer Science and Engineering, University of North Texas, Denton, Texas, United States

Corresponding Author: Samuel F Atkinson
Email address: atkinson@unt.edu

Background. The primary aim of the study reported here was to determine the effectiveness of utilizing local spatial variations in environmental data to uncover the statistical relationships between West Nile Virus (WNV) risk and environmental factors. Because least squares regression methods do not account for spatial autocorrelation and non-stationarity of the type of spatial data analyzed for studies that explore the relationship between WNV and environmental determinants, we hypothesized that a geographically weighted regression model would help us better understand how environmental factors are related to WNV risk patterns without the confounding effects of spatial non-stationarity.

Methods. We examined commonly mapped environmental factors using both ordinary least squares regression (LSR) and geographically weighted regression (GWR). Both types of models were applied to examine the relationship between WNV-infected dead bird counts and various environmental factors for those locations. The goal was to determine which approach yielded a better predictive model.

Results. LSR efforts lead to identifying three environmental variables that were statistically significantly related to WNV infected dead birds (adjusted R^2=0.61): stream density, road density, and land surface temperature. GWR efforts increased the explanatory value of these three environmental variables with better spatial precision (adjusted R^2 = 0.71).

Conclusions. The spatial granularity resulting from the geographically weighted approach provides a better understanding of how environmental spatial heterogeneity is related to WNV risk as implied by WNV infected dead birds, which should allow improved planning of public health management strategies.

A comparison of least squares regression and geographically weighted regression modeling of West Nile virus risk based on environmental parameters

Abhishek K. Kala[1], Chetan Tiwari[2], Armin Mikler[3], Samuel F. Atkinson[1]

[1]Advanced Environmental Research Institute, and Department of Biological Sciences, University of North Texas, Denton, Texas, USA
[2]Department of Geography, University of North Texas, Denton, Texas, USA
[3]Department of Computer Science and Engineering, University of North Texas, Denton, Texas, USA

Corresponding Author:
Samuel Atkinson
Advanced Environmental Research Institute, University of North Texas, 1155 Union Circle #310559, Denton, TX 76201, USA.
atkinson@unt.edu

Abstract

Background. The primary aim of the study reported here was to determine the effectiveness of utilizing local spatial variations in environmental data to uncover the statistical relationships between West Nile Virus (WNV) risk and environmental factors. Because least squares regression methods do not account for spatial autocorrelation and non-stationarity of the type of spatial data analyzed for studies that explore the relationship between WNV and environmental determinants, we hypothesized that a geographically weighted regression model would help us better understand how environmental factors are related to WNV risk patterns without the confounding effects of spatial non-stationarity.

Methods. We examined commonly mapped environmental factors using both ordinary least squares regression (LSR) and geographically weighted regression (GWR). Both types of models were applied to examine the relationship between WNV-infected dead bird counts and various environmental factors for those locations. The goal was to determine which approach yielded a better predictive model.

Results. LSR efforts lead to identifying three environmental variables that were statistically significantly related to WNV infected dead birds (adjusted R^2=0.61): stream density, road density, and land surface temperature. GWR efforts increased the explanatory value of these three environmental variables with better spatial precision (adjusted R^2 = 0.71).

Conclusions. The spatial granularity resulting from the geographically weighted approach provides a better understanding of how environmental spatial heterogeneity is related to WNV risk as implied by WNV infected dead birds, which should allow improved planning of public health management strategies.

39

40 Introduction

41 West Nile Virus (WNV) is a vector-borne disease that was first detected in the United
42 States in 1999 (Nash et al. 2001). Within a few years the virus had spread across the North
43 American continent (Hayes et al. 2005). WNV has had important environmental and human
44 impacts, including a decline in numerous bird species (CDC) and increased morbidity and
45 mortality among humans. This has also resulted in increased economic burdens due to initial
46 acute health care needs of infected individuals and subsequent long-term costs associates with
47 infection, estimated at approximately $56 million per year between 1999 and 2012 (Barrett
48 2014). Because that study indicated how difficult predicting and planning for WNV outbreaks
49 was, we became interested in developing a spatially explicit model using environmental factors
50 in an attempt to improve WNV risk predictions.

51 There are two important considerations that should typically be examined when
52 developing spatially explicit environmental disease risk models (Miller 2012). The first should
53 be an examination of potential spatial autocorrelation (the degree to which a set of spatial
54 features and their associated data values tend to be clustered together in space). This involves
55 accounting for whether environmental factors and the corresponding disease rates in
56 geographically proximate areas are more or less clustered together than they are in
57 geographically distant areas. Second, data non-stationarity (changing means, variances and
58 covariances in data across space) should be investigated and controlled when necessary
59 (Fotheringham 2009a; Miller 2012). Geographically weighted regression (GWR) can be used for
60 these two considerations and can often produce improved models that enable better spatial
61 inference and prediction. Recent studies have applied GWR modeling to drug-resistant

tuberculosis versus risk factors (Liu et al. 2011); environmental factors versus typhoid fever (Dewan et al. 2013); local climate and population distribution versus hand, foot, and mouth disease (Hu et al. 2012); and environmental factors and tick-borne disease (Atkinson et al. 2012; Atkinson et al. 2014; Wimberly et al. 2008a; Wimberly et al. 2008b), all showing that predictor variables varied spatially across large geographic regions, implying that the results for such studies may be improved using GWR.

The spatially explicit model that is discussed in this paper uses GWR to account for spatial heterogeneity for two reasons: (a) WNV disease risk observed across space may be related to similar environmental variables that increase vector habitat suitability and (b) environmental variables that influence WNV risk are not typically uniformly distributed across geographic space. Although many epidemiological models of WNV risk have been developed, it appears that there has been little research to explicitly examine techniques that account for spatial heterogeneity. Most models assume that the impact of various environmental factors are constant across the study region, which is unrealistic as larger areas display substantial variations in distribution of environmental, socio-economic, and demographic conditions (Goovaerts 2008).

Due to the unavailability of reliable and complete data, developing models of WNV risk pose additional challenges. Human case data is lacking due to issues of under-reporting and limited surveillance. Our alternative strategy was to assess WNV infected dead bird counts as a surrogate measure of human risk because “infection rates” in dead birds can be more precise because of the genetic markers tested in dead birds may be more reliable than case data and/or surveillance data. Additionally, others have also used mosquito habitat suitability as a surrogate for estimating WNV risk for human infection (Cooke et al. 2006). For our study, we followed a similar approach and used a model of mosquito habitat suitability condition as a predictor of the

spatial distributions of infected birds, which in turn can be used to estimate WNV disease risk among human populations. Further, because the environmental variables considered in this study are known to vary across space, we account for spatial autocorrelation and non-stationarity using GWR following the approach of (DeGroote et al. 2008) in order to improve the predictability of a model.

WNV transmission and risk factors

The WNV transmission cycle was an important component of our modeling efforts. The first step in the WNV transmission cycle primarily occurs when a competent female mosquito vector bites an infected bird reservoir host, which in turn results in the virus being transmitted to the mosquito (Blair 2009). This occurs when the female mosquito is seeking a blood meal to obtain nutrients necessary for egg development. After taking an infectious blood meal, a mosquito may pick up a permanent infection. The infected mosquito now has the potential to transmit the virus to another bird or animal when it feeds again. Once infected, birds may fly to other locations where the virus can be transmitted to susceptible mosquitoes. Subsequently, the disease may be transmitted by infected mosquitoes to humans or other mammals that act as incidental hosts. Dead birds found to be infected with WNV are often the primary indicators for presence of the disease in a geographic region and have proven to be useful for disease prediction modeling and identifying areas for human infection risk (Cooke et al. 2006; Ruiz et al. 2004; Valiakos et al. 2014). This relationship allows an assumption of a positive correlation between infected dead birds and WNV risk. Since the New York outbreak in 1999, WNV has been recovered from 26 mosquito species in North America, including *Culex. pipiens, Culex. salinarius, Culex. restuans, Ochlerotatus canadensis, Oc. japonicus, Aedes vexans*, and *Culiseta*

melanura (CDC 2000; Control & Prevention 2001). Results from a study (Goddard et al. 2002) assesses the vector competence of California mosquitoes. The results indicate that mosquitoes in the genus *Culex* (*Cx*.)are the principal hosts of WNV in California. The study also analyzed that on the basis of vector competence and host-feeding patterns, *Cx. tarsalis* may be the principal vector in rural agricultural ecosystems; and *Cx. pipiens* complex and *Cx. stigmatosoma* as important vectors in urban settings.

Vector and pathogen reservoirs overlap when certain environmental conditions are present (Rochlin et al. 2011). Table 1 provides an overview of the environmental conditions that are associated with WNV transmission, which were utilized for our research. These include characteristics of a place such as the mosquito species habitat: climatic conditions, topography and land use/land cover classes such as vegetation, water, and urbanized areas. Spectral indices acquired from satellite imagery provide information about environmental characteristics like temperature, vegetation cover, and moisture (Rodgers & Mather 2006). Liu and Weng (Liu & Weng 2012) in a study on WNV risk in southern California found that one of the main factors contributing to the WNV propagation included land surface temperature. They related higher temperature to viral replication in mosquitoes for WNV to be disseminated throughout the year. The results also show that areas with lower elevations tended to be more susceptible to WNV invasion as mosquito population propagates in the plain habitats with warmer temperatures compared to areas with higher elevation that have lower temperatures.

Table 1 approximately here

Statistical Considerations

Miller (2012) suggests that a 'global' model is the one that assumes that the parameters (commonly mean and variance) of some process are constant across geographic space

(commonly mean and variance), typically referred to as the spatial stationarity of a process. Miller suggests that in the case these parameters vary across geographic space (spatial heterogeneity), then such models may lead to inaccurate predictions and subsequent problems for decision-making. In an ecological context, spatial heterogeneity usually results from the interaction of various environmental processes that operate at different scales (Legendre 1993). Fotheringham (Fotheringham 2009b) used local statistics for linking the concepts of spatial autocorrelation and heterogeneity that are deemed important when developing spatial models. Local statistics disaggregate a global mean value into locally computed values for each spatial unit. It is based on a conceptualization of Tobler's first law in Geography (Tobler 1970) that specifies that "everything is related to everything else, but near things are more related than distant things." Spatial autocorrelation is a commonly used measure of the degree of spatial heterogeneity.

GWR is a local regression method that can be used for diagnosing spatial heterogeneity between dependent and explanatory variables over space (Fotheringham et al. 2003). It is performed within local windows centered on the nodes of a regular grid. Each observation within the local window is weighted based on its proximity to the center of that window. This approach has several advantages: it avoids abrupt changes in the local statistics computed for adjacent windows, helps visualize spatial variability within the geographic entity, and allows analysis of regionally aggregated data (Goovaerts 2008). A model's predictive ability, particularly in ecological modeling, is influenced not only by the strength of relationships between the species and its environment, but whether the model recognizes if the relationships are operating at multiple spatial scales. GWR provides a framework for exploring scale-dependent effects. It tests

153 the effect on a model's predictive ability by systematically increasing the size of the local
154 window (Miller 2012).

155 GWR can be used for mapping the spatial distribution of a model's coefficient values in
156 order to identify potential missing variables or to suggest other underlying factors associated
157 with the observed non-stationarity (Miller 2012). GWR is also useful for exploratory data
158 analysis and visualization; for example Kupfer and Farris (Kupfer & Farris 2007) used a 'leave-
159 one-out' (jackknifing) methodology to compare residuals from GWR and ordinary least squares
160 regression. They found that GWR often had more accurate predictions for sites that were
161 difficult to predict (where both models had overall higher residuals), which is why we used a
162 GWR framework for explicitly modeling the spatial relationships between WNV and its
163 environmental risk factors.

164

165

166

167 ## Materials & Methods

168 Study area

169 The model was built for the state of California, which was the national epicenter of WNV
170 activity in 2004 and 2005 (Jean et al. 2007). WNV was first detected there in July 2003 (Reisen
171 et al. 2004). It is the third largest state by area in the United States and is made up of 58 counties.
172 California has the largest population in the U.S., but it is unevenly distributed across the state.
173 The state also has a variable landscape with a large valley in the middle, bounded by mountain
174 ranges.

175

176 Environmental factors and data sources

177 Our model utilized various environmental factors (Table 1) that have been suggested as
178 descriptive in local WNV risk distribution: surface slope, density of roads, density of streams,
179 monthly mean temperature, monthly mean evapotranspiration, and land cover classes like
180 vegetation, developed land, cultivated land, and open surface water. All environmental
181 parameters except roads and streams (Table 2) were acquired in grid format and resampled to
182 120 meter resolution as suggested by Cooke et al. (2006). Data resampling was done using the
183 resampling tool available in ArcGIS software. The modeling method utilized in this study was
184 based on analyzing data in raster format, and therefore road and stream vector data were
185 converted to raster format using the 'Kernel Density Estimation' tool in ArcGIS to create road
186 density and stream density grid files. The tool assumes a Gaussian distribution and thus assigns
187 more importance towards the center of kernel in comparison to the features that are further apart.

188 Various dynamic environmental data including Normalized Difference Vegetation Index
189 (NDVI), Land Surface Temperature (LST), and Evapotranspiration (ET) were downloaded from
190 the Moderate Resolution Imaging Spectroradiometer (MODIS) toolbox incorporated in
191 ArcGIS®. The Land Surface Temperature tool accesses MOD11-A1, the daily averaged LST
192 product. The MOD11 product uses the algorithm which is optimally used to separate ranges of
193 atmospheric column water vapor and lower boundary air surface temperatures into tractable sub-
194 ranges. The NDVI is calculated according to the formula NDVI = (NIR - VIS)/ (NIR + VIS)
195 where NIR is the near-infrared radiance and VIS is observed radiation in the visible spectrum.
196 NDVI data is available from either satellite with MODIS (Aqua or Terra) as a monthly average.
197 The time lag between the hatching of a mosquito egg to an adult mosquito taking blood meals
198 and becoming infected with WNV to the subsequent infection of a human and the appearance of

WNV disease symptoms was taken into account and therefore environmental data used for this study was taken for the month of July, the month prior to peak WNV human incidence cases (Campbell et al. 2002).

Table 2 approximately here

Least Squares Regression (LSR) modeling

WNV disease annual incidence rate (cases per 100,000 populations) was used as the measure of disease severity in this study. Annual WNV-infected dead birds sentinel data, averaged for 2004-2010, was used as a surrogate of WNV risk and was the dependent variable for modeling purposes in this study because several other studies (Chaintoutis et al. 2014; Eidson et al. 2001a; Eidson et al. 2001b; Eidson et al. 2001c; Guptill et al. 2003; Johnson et al. 2006; Mostashari et al. 2003; Nielsen & Reisen 2007; Patnaik et al. 2007; Ruiz et al. 2004) have suggested links between infected dead birds and WNV human infection rates. Since wild birds are the primary reservoir hosts for WNV and indicator of human infection risk, we utilized this association to develop the disease prediction model. We determined the utility of this relationship by correlating the dead birds data with the human incidence rate ($r^2 = 0.409$). While infected dead bird counts only explain about 40% of reported human cases in California, it is a highly significant predictor ($p=0.01$). Hence we used a dead bird model, with infected dead birds as a dependent variable, to assess WNV risk among human population.

Interpretations of ordinary Least Squares Regression (LSR) model performance were based on assessing multi-collinearity, robust probability, adjusted R^2, and Akaike's information criterion (AIC) (Akaike 1974). Multi-collinearity was assessed through the variance inflation factor (VIF) statistic, which measures redundancy among explanatory variables. Explanatory variables associated with VIF values larger than about 7.5 indicate that these variables are

223 providing similar information, and they were removed one at a time from the model based on
224 VIF value until the model became unbiased. Robust probability indicates the statistically
225 significant variables that are important to the regression model. Examining VIF values and
226 robust probability, we ran and re-ran LSR models until narrowing down to non-redundant and
227 significant variables: land surface temperature; stream density, and; road density. Akaike's
228 information criterion (AIC) was then used to determine the best LSR model.
229 The next step was to explore GWR models that might better explain the variation in
230 infected dead bird counts based upon environmental data. Spatial autocorrelation (Global
231 Moran's I) was utilized to assess whether the environmental factors exhibited a random spatial
232 pattern (Goodchild 1986), and where adequate models have a random distribution of the
233 residuals (Mitchell 2005).
234
235 Geographically Weighted Regression (GWR) modeling
236 Under conditions of non-stationarity in LSR modeling, geographically weighted
237 regression (GWR) was explored to potentially improve modeling results. The same explanatory
238 variables that were used in LSR modeling were used to run GWR rather than starting with the
239 full global set of parameters so as to avoid introducing "improvement" that could not be
240 attributed solely to which modeling approach was applied. In other words, if GWR modeling was
241 not applied to the same variables as LSR modeling, but yielded better results, we would not
242 know if the improvement was due to the modeling approach or the environmental data that was
243 used to build each model.
244 Once key environmental factors were identified during LSR modeling, we proceeded to
245 explore the spatial variability of local regression coefficients to determine whether the

underlying process exhibited spatial heterogeneity (Fotheringham et al. 2003). A GWR local model was applied to analyze how the relationship between infected dead bird counts and environmental factors changed from one county to another. Unlike conventional LSR regression modeling, which produces a single regression equation to summarize global relationships among the independent and dependent variables, GWR detects spatial variation within relationships in a model and produces information useful for exploring and interpreting spatial non-stationarity (Fotheringham et al. 2003).

A spatial kernel was used to provide geographic weighting for the local window centered on the grid nodes used in our model. There are two possible categories of spatial kernels: fixed/adaptive and bandwidth, which is a key coefficient that controls the size of the kernel (Akaike 1974). These kernels tend to be Gaussian or Gaussian-like which implies that distant samples are weighed lesser than the proximal ones. There are three potential bandwidth approaches: Akaike information criterion (AIC), cross validation (CV), and bandwidth parameter. For our GWR model, the AIC approach was chosen because the distribution of infected dead birds was not consistent in the study area. The following settings were used in ArcGIS GWR: Bandwidth method = AIC and Kernel type = Adaptive.

Finally, we examined independency and normality of residuals, to evaluate the fit of the model. Local collinearity, the square root of the largest eigenvalue divided by the smallest eigenvalue, of our GWR model was also assessed but no data points were removed as they compromised model diagnostics. The adjusted coefficient of determination (Adjusted R^2) was used for comparing LSR and GWR models to determine which approach would provide a better understanding of the relationship between environmental conditions and West Nile Virus risk (Fotheringham et al. 2003).

Results

LSR modeling identified land surface temperature (VIF = 1.046), stream density (VIF = 1.177), and road density (VIF = 1.143) as statistically significant ($p < 0.05$) variables related to WNV risk:

$$\text{WNV risk} = -75.87 + 595.60\,(RD) + 1.89\,(LST) - 146.89\,(SD) \quad (1)$$

Where:

WNV risk = average infected dead bird count

RD = road density

LST = land surface temperature

SD = stream density

The histogram of the LSR model's residuals approximates that of a normal curve, with a non-significant (0.134, $p<0.05$) Jarque-Bera statistic (Jarque & Bera 1980), and the Moran's I Index Z-score (1.23) all imply that the model is unbiased and significantly different than random.

However, the Koenker statistic (0.000007*, $p<0.05$) confirmed non-stationarity in the LSR model indicating that there is not a consistent relationship between the explanatory variables and WNV risk across the study area. Further, the presence of mild heteroskedasticity was noted in the LSR model. We conclude that the LSR model is stable but non-stationary, suggesting that proceeding with GWR model was warranted.

The GWR model in this study was implemented using the following algorithm:

$$\text{WNV risk}_{(i)} = \beta_{i0} + \beta_{(i1)} RD_{(i)} + \beta_{(i2)} LST_{(i)} - \beta_{(i3)} SD_{(i)} + \varepsilon_{(i)} \quad (2)$$

Where β coefficients are county (i) specific, and RD is road density, LST is land surface temperature, and SD is stream density.

Comparing the fit of the global LSR model (assumes homogeneity of variables across space) and local GWR model (makes no assumption of homogeneity), we found that the global LSR adjusted R^2 was 0.61 (R^2 was 0.66, $P<0.05$, Fig. 1) with analysis run on all 58 counties. The local GWR adjusted R^2 was 0.71 (R^2 is 0.75, $p<0.05$, Fig. 2) with a bandwidth of 54, which suggests that there has been some improvement by using a local modeling approach. Our preferred measure of model fit, AIC, gave a value of 567.7 for the global model and 551.4 for the local model. The difference of 16.3 is relatively strong evidence of an improvement in the model fit to the data. Further, the problem of hetroskedasticity that was noted in the OLS model was not observed in the GWR model.

We also tested the results using different bandwidth parameter. Several iterations were run but it was observed that although a smaller band-width criterion gave an improved combination of AICc and adjusted R^2 values, it also compromised the model diagnostics by introducing local collinearity and thus instability in the model. Addressing local collinearity by removing the Counties having condition number greater than 30 affected the model's overall results. Thus, it is better to have a larger band-width rather than violating model assumptions and to avoid the unstable prediction (Charlton & Fotheringham ; Nakaya 2014).

Figure 1 approximately here

Figure 2 approximately here

316 Mapping the values of the standardized residual across California (Fig. 3a) provides a
317 representation of: (a) areas with unusually high or low residuals and (b) whether the residuals
318 were spatially autocorrelated. Counties with excessively large positive residuals would under-
319 predict WNV risk, and counties with excessively large negative residuals would over-predict
320 WNV risk. The spatial autocorrelation of GWR residuals for our model resulted in a Moran's I
321 value of -0.11 (p=0.18), implying little evidence of any autocorrelation in them.

322 **Figure 3 approximately here**

323 Local coefficient estimates for significant factors were mapped using quantile
324 classification method. Figure 3b shows the variation in the model's coefficient estimates for the
325 land surface temperature (LST) variable. The map for the local coefficients reveals that the
326 influence of this variable in the model varies considerably over California, with a strong north-
327 south direction. The range of the local coefficient is from 1.26 for the northernmost counties to
328 3.06 for the southernmost counties – evidence that points to heterogeneity in the model structure
329 within California. The global coefficient and all the local coefficients for this variable are
330 positive – there is agreement between the two models on the direction of the influence of this
331 variable. Figure 3c shows a similar distribution in north-south direction of positive road density
332 coefficient. Figure 3d reveals the opposite for stream density coefficients, with larger values in
333 the north and smaller values in south. Contrary to our initial thoughts, stream density
334 demonstrated a negative relation to disease risk. This may reflect that flowing water is normally
335 not suitable for larval development of the various species of mosquitos that commonly transmit
336 WNV or that rasterizing the stream database into stream density introduces a component that is
337 not yet fully understood.

Our best ordinary least squares model, the global LSR model (equation 1) produced an adjusted R^2 of 0.61 ($p<0.05$) with a corresponding corrected AIC of 567.70. Utilizing the same environmental variables, our best local GWR model (equation 2) produced an adjusted R^2 of 0.71 ($p<0.05$) with a corresponding corrected AIC of 551.4. A 16 point decrease in the AIC and approximately 16% improvement in the model performance suggest that incorporating spatial data improves the predictive ability of WNV risk.

Conclusions

One of the frequent technical issues in modeling disease risk is to incorporate local rather than global associations in these models (Foley et al. 2009). In spatial regression models, a global model can be used to examine the relationship between disease risk and potential explanatory factors which are based on the assumption that the relationship is a stationary spatial process (Miller 2012). For a small and homogenous region of interest, it is reasonable to assume that the explanatory factors would not change significantly across the region, and the relationship between WNV risk and the potential factors would also be unchanged. However, important variables such as topography, climate, and population distribution change greatly when it comes to a large region like California with an area of over 163,000 square miles. California is geographically diverse and is equally varied in its range of climates with several climatic sub-regions recognized. It would be unexpected to find that the spatial stationarity assumption holds in such a large area.

The distinct north-south pattern revealed in our study could be attributable to typical latitudinal expressions of temperature and precipitation, especially since California has a north-south length of 1,350 kilometers. This environmental pattern is also a likely contributor to the

361 distribution of different mosquito species in the United Stated, especially notable in its
362 manifestation in California. A recent report (CDC 2013) shows that while *Cx. Tarsalis* is
363 distributed throughout California, *Cx. pipiens* is a more important mosquito vector in northern
364 California, while *Cx. quinquefasciatus* is more important in southern California. While WNV
365 can be found in a wide variety of ecosystems, the north-south pattern of infected birds detected
366 in this study may be expressed more noticeably in California due to the north-south differences
367 in mosquito species distributions as observed in the Centers for Disease Control report.

368 Our results concur that understanding WNV risks is improved when considering spatial
369 heterogeneity of the variables that affect the risk (Beck et al. 1994). Besides improving
370 prediction accuracy, spatial heterogeneity can also provide insights into the underlying
371 ecological processes controlling the distributions of vector populations and zoonotic pathogens
372 (Wimberly et al. 2008a) because GWR models consider spatial heterogeneity by separating the
373 large heterogeneous region into smaller, more homogeneous local regions. Fotheringham
374 (Fotheringham 2009a), stated that an advantage of using GWR is that it accounts for much of the
375 spatial autocorrelation in the residuals that is usually found in global modeling. Further, it is
376 possible that a variable that is insignificant at the global level might be important locally.

377 There are several limitations of this study. First, it is assumed that factors suitable for
378 mosquito habitat increase the likelihood of WNV spread in human populations. On the surface
379 this seems to be reasonably apparent; however, we do not have specific evidence that this is true.
380 Second, it is also assumed that the probability of human infection is higher in counties with
381 multiple confirmed WNV bird cases, another reasonable conjecture with several references in the
382 literature, but without direct confirmation. Potential problems with this assumption include
383 varying human population density (e.g. two areas with the same number of infected dead birds

reported but one area's human population density is substantially different than the other), variations in level of public concern (as reports of infected dead birds increase, more people begin looking for dead birds), and resource availability might bias the reporting of dead birds (wealthy areas devote disproportionate resources to the issue). Thus, proper surveillance methods that take into consideration these limitations while collecting infected dead bird data will contribute to more meaningful results. Third, our approach assumes that people are infected within the county of their residence, ignoring the possibility of contracting an infection while traveling outside the county limits. Lastly, road density could also be correlated with dead bird surveillance effort and might be a potential bias for reporting dead birds. We recognize that if these assumptions do not hold, modeling WNV risks based on infected dead birds may yield biased results. However, if the assumptions do hold, the local modeling approaches should improve predictions of WNV risks.

The research described in this paper suggests that a spatially explicit local model using GWR approaches to adjust for spatial autocorrelation and non-stationarity can yield improved predictions compared to ordinary LSR modeling of WNV risk. A spatially explicit modeling technique may be useful in policy-making and decision-making depending on the granularity and resolution of available data. Identifying the spatial variations in relationships by estimating local regression parameters allows the spatial distribution and interaction of predictor variables to be explored. Analyzing local variations in relationships provides those concerned with public health policy the ability to target resources and to achieve improved outcomes through location-specific activities (Comber et al. 2011) because spatial heterogeneity can improve predictions by capturing geographic shifts in the ecological drivers (Wimberly et al. 2008a). While environmental data used in this research were of fine resolution, WNV disease human incidence

data and infected dead bird data that is used is available only at a coarser county scale. We had to assume that aggregating the environmental data up to the county adequately represented the environmental conditions presented in the county, but we knew that data aggregation was likely to introduce some uncertainty into the model. The dead bird model applied in this study can be used for better understanding of WNV risk and the techniques used could be replicated at finer spatial scales thus leading to better intervention efforts.

In summary, WNV, a globally emerging infectious disease, was found to be heterogeneously related to environmental factors at the county level throughout California during the time that our data were collected. Our findings may assist those conducting risk assessments for WNV transmission in local areas by helping local public health entities allocate resources and improve preparedness for an outbreak according to region-specific conditions.

Acknowledgments

The authors wish to thank the Center for Computational Epidemiology and Response Analysis, the Center for Remote Sensing, and the Advanced Environmental Research Institute at the University of North Texas for technical advice and access to software and hardware to undertake this project.

References

Akaike H. 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* 19:716-723.

Atkinson SF, Sarkar S, Aviña A, Schuermann JA, and Williamson P. 2012. Modelling spatial concordance between Rocky Mountain spotted fever disease incidence and habitat probability of its vector Dermacentor variabilis (American dog tick). *Geospatial health* 7:91-100.

Atkinson SF, Sarkar S, Aviña A, Schuermann JA, and Williamson P. 2014. A determination of the spatial concordance between Lyme disease incidence and habitat probability of its primary vector Ixodes scapularis (black-legged tick). *Geospatial health* 9:203-212.

Barrett AD. 2014. Economic Burden of West Nile Virus in the United States. *The American journal of tropical medicine and hygiene* 90:389-390.

Beck LR, Rodriguez MH, Dister SW, Rodriguez AD, Rejmankova E, Ulloa A, Meza RA, Roberts DR, Paris JF, and Spanner MA. 1994. Remote sensing as a landscape epidemiologic tool to identify villages at high risk for malaria transmission. *The American journal of tropical medicine and hygiene* 51:271-280.

Blair C. 2009. Vector biology and West Nile Virus. In: Diamond M, ed. *West Nile Encephalitis Virus Infection: Viral Pathogenesis and the host immune response*: Springer, 45-61.

Campbell GL, Marfin AA, Lanciotti RS, and Gubler DJ. 2002. West nile virus. *The Lancet infectious diseases* 2:519-529.

CDC. Species of dead birds in which West Nile virus has been detected, United States, 1999-2012. *Available at http://www.cdc.gov/westnile/resources/pdfs/Bird%20Species%201999-2012.pdf* (accessed 01/15 2014).

CDC. 2000. Update: West Nile virus activity—Northeastern United States, 2000. *MMWR Morb Mortal Wkly Rep* 49::820–822.

CDC. 2013. West Nile Virus in the United States: Guidelines for Surveillance, Prevention, and Control. Fort Collins, Colorado: U.S. Department of Health and Human Services. p 69.

Chaintoutis SC, Dovas CI, Papanastassopoulou M, Gewehr S, Danis K, Beck C, Lecollinet S, Antalis V, Kalaitzopoulou S, and Panagiotopoulos T. 2014. Evaluation of a West Nile virus surveillance and early warning system in Greece, based on domestic pigeons. *Comparative immunology, microbiology and infectious diseases* 37:131-141.

458 Charlton M, and Fotheringham A. nd Geographically Weighted Regression: A Tutorial on Using
459 GWR in ArcGIS 9.3. National Centre for Geocomputation. National University of
460 Ireland.

461 Comber AJ, Brunsdon C, and Radburn R. 2011. A spatial analysis of variations in health access:
462 linking geography, socio-economic status and access perceptions. *International Journal*
463 *of Health Geographics* 10:44-55.

464 Control CfD, and Prevention. 2001. West Nile virus activity--eastern United States, 2001.
465 *MMWR Morbidity and mortality weekly report* 50:617.

466 Cooke WH, Grala K, and Wallis RC. 2006. Avian GIS models signal human risk for West Nile
467 virus in Mississippi. *International Journal of Health Geographics* 5:36.

468 DeGroote J, Sugumaran R, Brend S, Tucker B, and Bartholomay L. 2008. Landscape,
469 demographic, entomological, and climatic associations with human disease incidence of
470 West Nile virus in the state of Iowa, USA. *International Journal of Health Geographics*
471 7:19.

472 Dewan AM, Corner R, Hashizume M, and Ongee ET. 2013. Typhoid fever and its association
473 with environmental factors in the Dhaka Metropolitan Area of Bangladesh: a spatial and
474 time-series approach. *PLoS neglected tropical diseases* 7:e1998.

475 Eidson M, Komar N, Sorhage F, Nelson R, Talbot T, Mostashari F, McLean R, and Group
476 WNVAMS. 2001a. Crow deaths as a sentinel surveillance system for West Nile virus in
477 the northeastern United States, 1999. *Emerging infectious diseases* 7:615.

478 Eidson M, Kramer L, Stone W, Hagiwara Y, Schmit K, and Team NYSWNVAS. 2001b. Dead
479 bird surveillance as an early warning system for West Nile virus. *Emerging infectious*
480 *diseases* 7:631.

481 Eidson M, Miller J, Kramer L, Cherry B, Hagiwara Y, and Group WNVBMA. 2001c. Dead crow
482 densities and human cases of West Nile virus, New York State, 2000. *Emerging*
483 *infectious diseases* 7:662.

484 Foley R, Charlton MC, and Fotheringham AS. 2009. GIS in Health and Social Care Planning.

485 Fotheringham AS. 2009a. Geographically weighted regression. *The Sage handbook of spatial*
486 *analysis*:243-254.

Fotheringham AS. 2009b. “The problem of spatial autocorrelation” and local spatial statistics. *Geographical analysis* 41:398-403.

Fotheringham AS, Brunsdon C, and Charlton M. 2003. *Geographically weighted regression: the analysis of spatially varying relationships*: John Wiley & Sons.

Goddard LB, Roth AE, Reisen WK, and Scott TW. 2002. Vector competence of California mosquitoes for West Nile virus. *Emerging infectious diseases* 8:1385-1391.

Goodchild MF. 1986. *Spatial autocorrelation*: Geo Books Norwich.

Goovaerts P. 2008. Geostatistical analysis of health data: State-of-the-art and perspectives. *geoENV VI–Geostatistics for Environmental Applications*: Springer, 3-22.

Guptill SC, Julian KG, Campbell GL, Price SD, and Marfin AA. 2003. Early-season avian deaths from West Nile virus as warnings of human infection. *Emerging infectious diseases* 9:483.

Hayes EB, Komar N, Nasci RS, Montgomery SP, O'Leary DR, and Campbell GL. 2005. Epidemiology and transmission dynamics of West Nile virus disease. *Emerging infectious diseases* 11:1167.

Hu M, Li Z, Wang J, Jia L, Liao Y, Lai S, Guo Y, Zhao D, and Yang W. 2012. Determinants of the incidence of hand, foot and mouth disease in China using geographically weighted regression models. *PloS one* 7:e38978.

Jarque CM, and Bera AK. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters* 6:255-259.

Jean CM, Honarmand S, Louie JK, and Glaser CA. 2007. Risk factors for West Nile virus neuroinvasive disease, California, 2005. *Emerging infectious diseases* 13:1918.

Johnson GD, Eidson M, Schmit K, Ellis A, and Kulldorff M. 2006. Geographic prediction of human onset of West Nile virus using dead crow clusters: an evaluation of year 2002 data in New York State. *American journal of epidemiology* 163:171-180.

Kupfer JA, and Farris CA. 2007. Incorporating spatial non-stationarity of regression coefficients into predictive vegetation models. *Landscape Ecology* 22:837-852.

514 Legendre P. 1993. Spatial autocorrelation: trouble or new paradigm? *Ecology* 74:1659-1673.

515 Liu H, and Weng Q. 2012. Environmental factors and risk areas of West Nile Virus in southern
516 California, 2007–2009. *Environmental Modeling & Assessment* 17:441-452.

517 Liu Y, Jiang S, Wang R, Li X, Yuan Z, Wang L, and Xue F. 2011. Spatial epidemiology and
518 spatial ecology study of worldwide drug-resistant tuberculosis. *Int J Health Geogr*
519 10:10.1186.

520 Miller JA. 2012. Species distribution models Spatial autocorrelation and non-stationarity.
521 *Progress in Physical Geography* 36:681-692.

522 Mitchell A. 2005. The ESRI Guide to GIS Analysis: Spatial Measurements and Statistics. Vol 2.
523 Redlands. CA: ESRI Press.

524 Mostashari F, Kulldorff M, Hartman JJ, Miller JR, and Kulasekera V. 2003. Dead bird clusters
525 as an early warning system for West Nile virus activity. *Emerging infectious diseases*
526 9:641.

527 Nakaya T. 2014. GWR4 user manual. *WWW document,*
528 *https://geodacenterasuedu/drupal_files/gwr/GWR4manualpdf*.

529 Nash D, Mostashari F, Fine A, Miller J, O'Leary D, Murray K, Huang A, Rosenberg A,
530 Greenberg A, and Sherman M. 2001. The outbreak of West Nile virus infection in the
531 New York City area in 1999. *New England Journal of Medicine* 344:1807-1814.

532 Nielsen CF, and Reisen WK. 2007. West Nile virus-infected dead corvids increase the risk of
533 infection in Culex mosquitoes (Diptera: Culicidae) in domestic landscapes. *Journal of*
534 *medical entomology* 44:1067-1073.

535 Patnaik JL, Juliusson L, and Vogt RL. 2007. Environmental predictors of human West Nile virus
536 infections, Colorado. *Emerging infectious diseases* 13:1788.

537 Reisen W, Lothrop H, Chiles R, Madon M, Cossen C, Woods L, Husted S, Kramer V, and
538 Edman J. 2004. West Nile virus in California. *Emerging infectious diseases* 10:1369.

539 Rochlin I, Turbow D, Gomez F, Ninivaggi DV, and Campbell SR. 2011. Predictive mapping of
540 human risk for West Nile virus (WNV) based on environmental and socioeconomic
541 factors. *PloS one* 6:e23280.

542 Rodgers SE, and Mather TN. 2006. Evaluating satellite sensor-derived indices for Lyme disease
543 risk prediction. *Journal of medical entomology* 43:337-343.

544 Ruiz MO, Tedesco C, McTighe TJ, Austin C, and Kitron U. 2004. Environmental and social
545 determinants of human risk during a West Nile virus outbreak in the greater Chicago
546 area, 2002. *International Journal of Health Geographics* 3:8.

547 Tobler WR. 1970. A computer movie simulating urban growth in the Detroit region. *Economic*
548 *geography*:234-240.

549 Valiakos G, Papaspyropoulos K, Giannakopoulos A, Birtsas P, Tsiodras S, Hutchings MR,
550 Spyrou V, Pervanidou D, Athanasiou LV, and Papadopoulos N. 2014. Use of Wild Bird
551 Surveillance, Human Case Data and GIS Spatial Analysis for Predicting Spatial
552 Distributions of West Nile Virus in Greece. *PloS one* 9:e96935.

553 Wimberly MC, Baer AD, and Yabsley MJ. 2008a. Enhanced spatial models for predicting the
554 geographic distributions of tick-borne pathogens. *International Journal of Health*
555 *Geographics* 7:15.

556 Wimberly MC, Hildreth MB, Boyte SP, Lindquist E, and Kightlinger L. 2008b. Ecological niche
557 of the 2003 west nile virus epidemic in the northern great plains of the United States.
558 *PloS one* 3:e3744.

559

Table 1(on next page)

Environmental conditions related to WNV transmission risk.

1

Factors studied	Relation to WNV	References
Streams	Sites for breeding and resting	Cooke et al. 2006; Curtis et al. 2014; Schurich et al. 2014
Temperature	Increases growth rate of vector, decreases egg development cycle and shortens extrinsic incubation period of vector	DeGroote et al. 2014; Kuehn 2012; Srivastava et al. 2001; Wimberly et al. 2008b
Surface slope	Water stagnation creating mosquito breeding ground	Cooke et al. 2006; Ozdenerol et al. 2008; Schurich et al. 2014; Srivastava et al. 2001
Cultivated land	Linkage between habitat used and human-commensal nature of WNV mosquito vectors	Kilpatrick 2011
Developed land	Linkage between habitat used and human-commensal nature of WNV mosquito vectors; warmer microclimates	Kilpatrick 2011
Roads	Sites for breeding and resting along roadsides	Cooke et al. 2006
Vegetation	Sites for breeding and resting.	Brownstein et al. 2002; Cooke et al. 2006; DeGroote et al. 2014; Ruiz et al. 2004a; Schurich et al. 2014; Srivastava et al. 2001
Evapotranspiration	Related to the amount of moisture that is related to mosquito abundance	Liu and Weng 2012; Trawinski and Mackay 2008

2

3

Table 2(on next page)

Data sources.

1

Data	Spatial resolution	Source
Elevation	10m	National Elevation Dataset (NED)
LST	1 Km	MODIS aboard the Terra and Aqua satellites
NDVI	250 m	MODIS aboard the Terra and Aqua satellites
Evapotranspiration (ET)	1 Km	MODIS aboard the Terra and Aqua satellites
Streams	available in vector format	U.S. bureau of reclamation
Roads	available in vector format	U.S. Census bureau
Cultivated land	30 m	National Land Cover Database
Developed land	30 m	National Land Cover Database
WNV infected dead birds count	County scale	U.S.G.S. National wildlife health center
WNV human incidence cases	County scale	U.S.G.S. National wildlife health center
Human population	County scale	U.S. Census bureau

2

Figure 1

Trendline plot for global LSR model (model: y = 0.6591x + 10.563; r2 = 0.66), dashed line ideal 1:1 relationship.

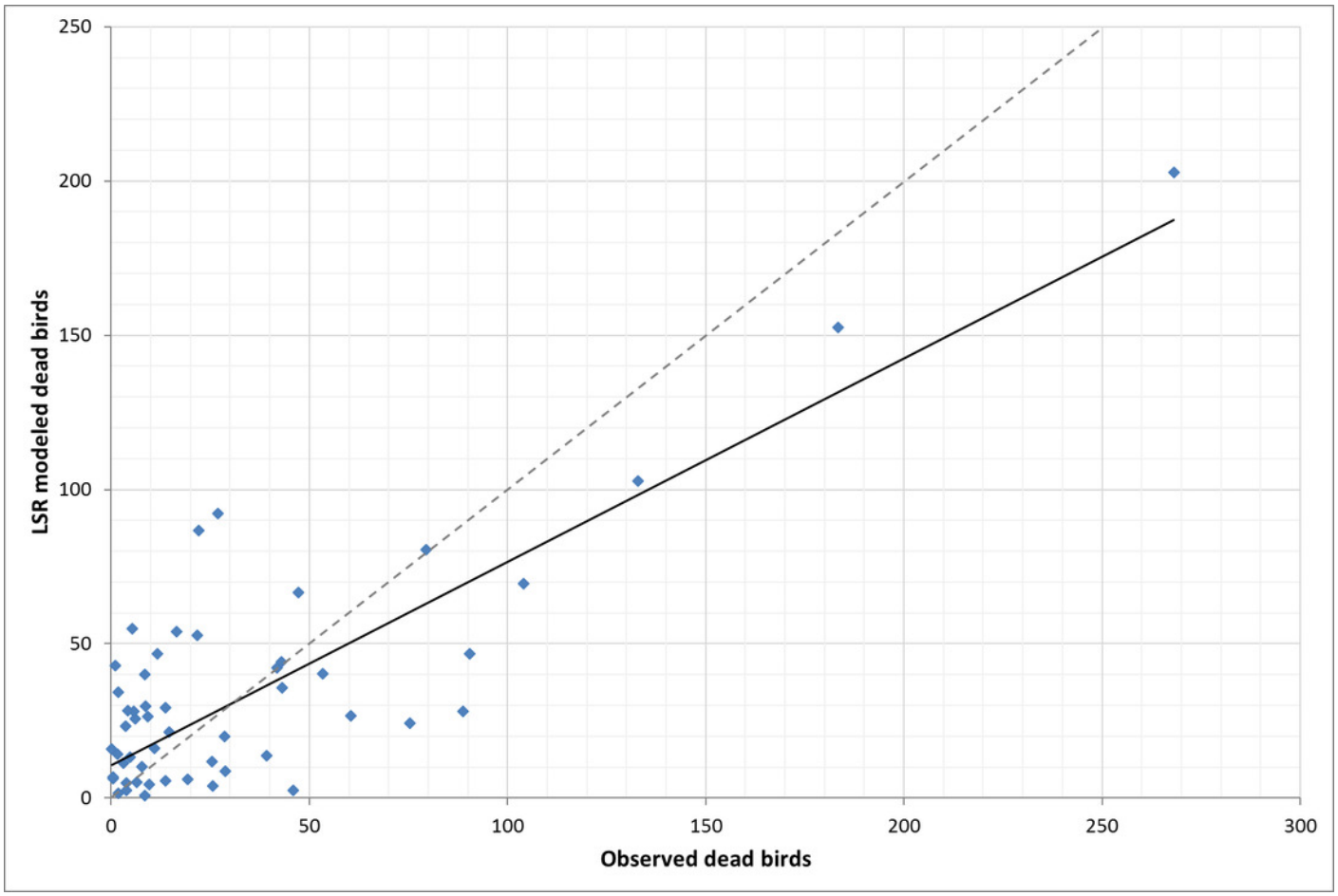

Figure 2

Trendline plot for local GWR model (model: y = 0.6911x + 10.259; r2 = 0.75), dashed line ideal 1:1 relationship.

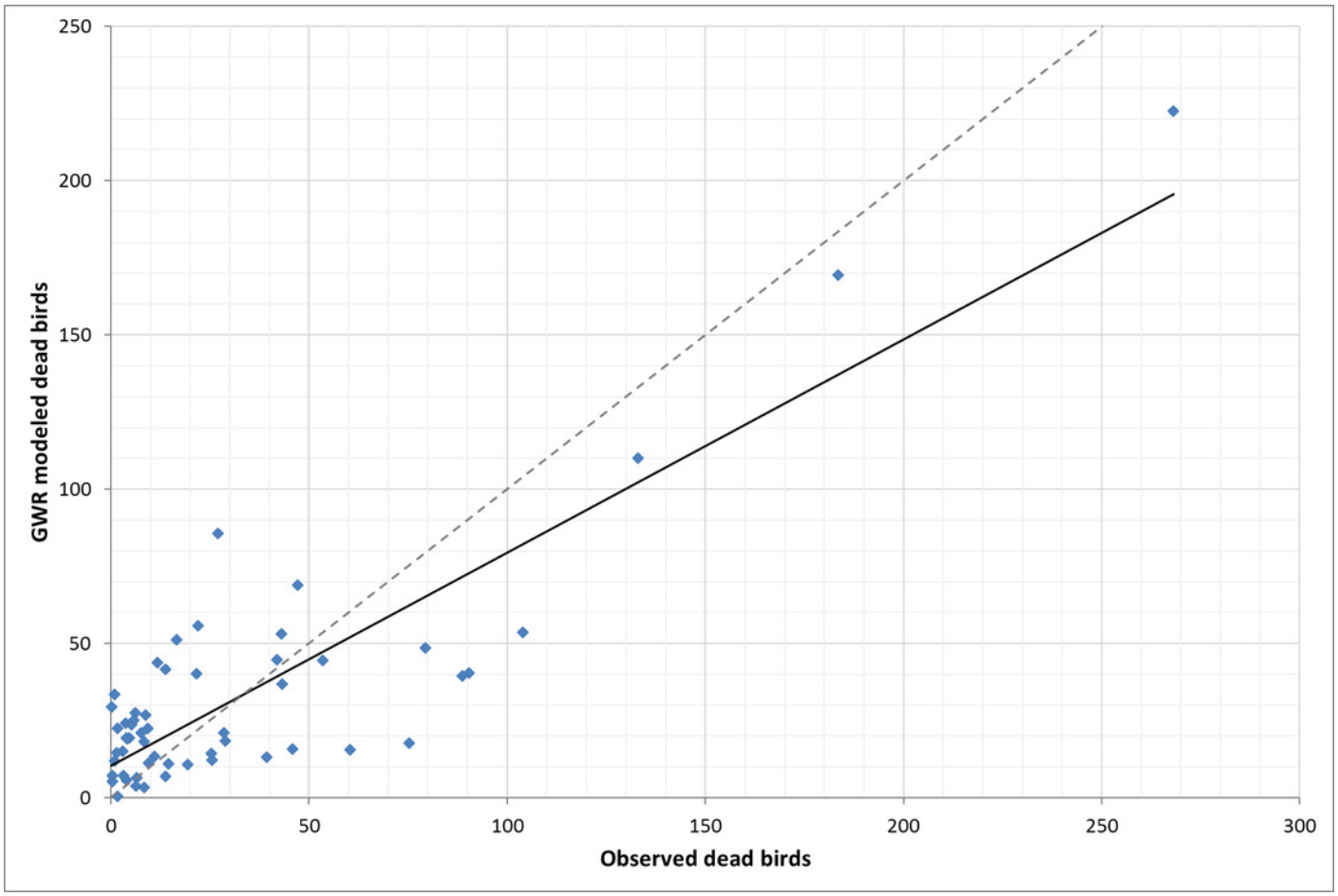

Figure 3

Spatial distribution of (a) standardized residuals; (b) land surface temperature coefficients; (c) road density coefficients; (d) stream coefficients.