MODULE 9: Spatial Statistics in Epidemiology and Public Health
Lecture 5: Slippery Slopes: Spatially Varying Associations

Jon Wakefield and **Lance Waller**

What are we doing?

- Alcohol
- Illegal drugs
- Violent crimes
- Regression
- Breaking the rules (what if the associations change in space?)

Spatially varying associations

Violence, alcohol, drugs
- Building the model
- Results

Conclusions

Acknowledgements and References

- Collaborators: Paul Gruenewald, Dennis Gorman, Li Zhu, Carol Gotway, and David Wheeler
- References:
  - Waller et al. (2008) Quantifying geographical associations between alcohol distribution and violence... *Stoch Environ Res Risk Assess* **21**: 573-588.
  - Wheeler and Caldor (2009) As assessment of coefficient accuracy...*J Geogr Systems* **9**: 573-588.
  - Wheeler and Waller (2009) Comparing spatially varying coefficient models... *J Geogr Systems* **11**: 1-22.
  - Finley (2011) Comparing spatially-varying coefficient models...*Methods in Ecology and Evolution* **2**: 143-154.

What do we want to do?

- Quantify associations between outcomes and covariates as observed in data.
- Adjust for spatial correlation (spatial regression) using a random intercept with a CAR prior.
- What if strength of association varies across space?
- Usually, we assume β is the same at every location, what if it varies (but is spatially correlated)?
- Can we have a *random slope*? Can we use CAR priors for that?

What about spatially varying associations?

- Fix it: Geographically weighted regression (GWR)
  - Fotheringham et al. (2002)
- Model it: Spatially varying coefficient (SVC) models
  - Leyland et al. (2000), Assuncao et al. (2003), Gelfand et al. (2003), Gamerman et al. (2003), Congdon (2003, 2006)

Our data for today

- Outcome: Rates (number of cases per person per year) of violent crimes (police/sheriff reports).
- Covariates: Alcohol distribution (licenses and sales), illegal drug arrests (police/sheriff reports).
- Potential confounders: Sociodemographics (census).
- Linked to common spatial framework (census tracts) via GIS.

Translation complications

- When are crime data like disease data?
  - Counts from small areas.
  - Per person "rate" of interest.
- When are crime data not like disease data?
  - Outcome not as "rare".
  - Police vs. medical records.
  - Residents not only ones at risk.

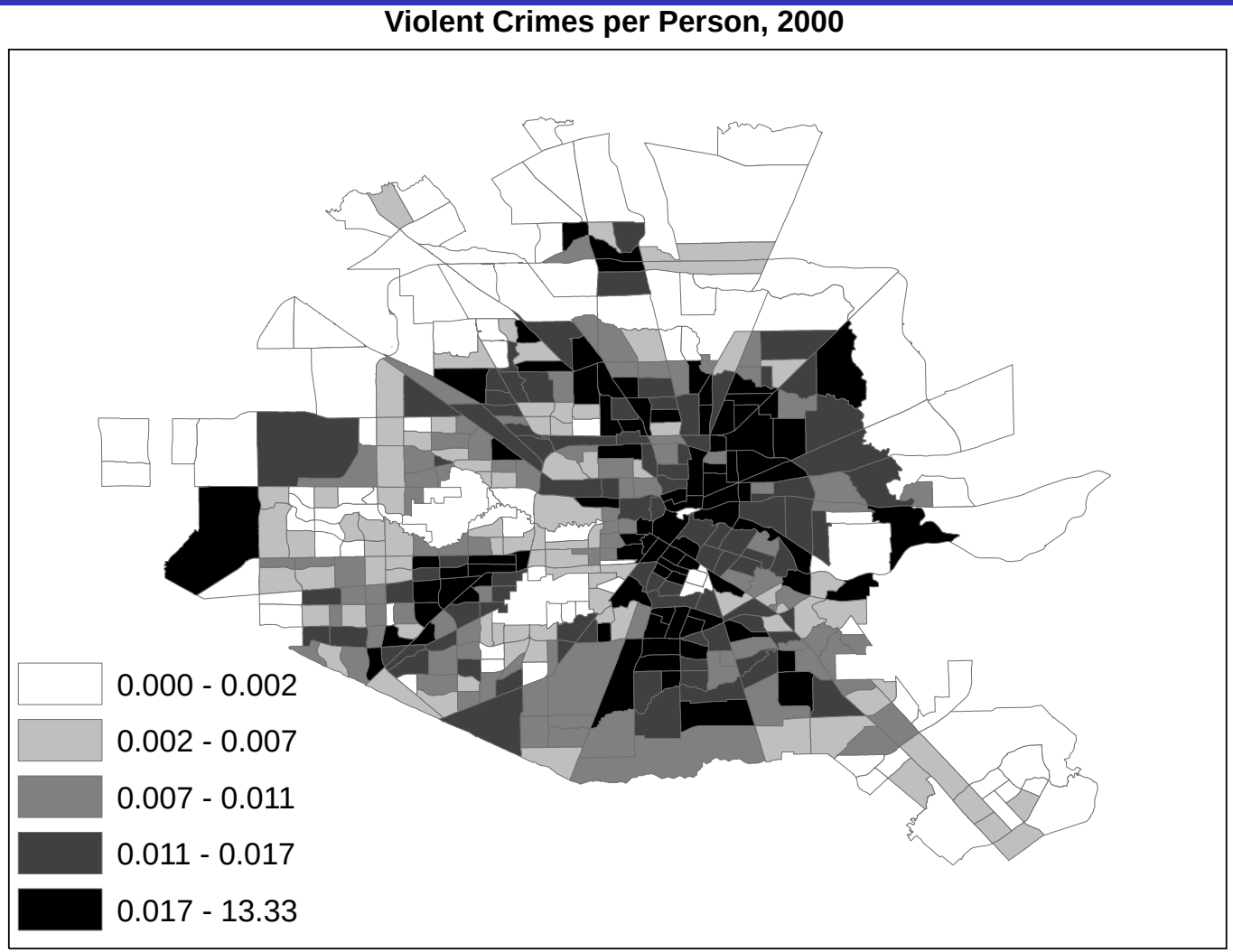
Background: Alcohol, drug arrests, and violent crime

- Large body of research showing association between alcohol distribution and the incidence of violence.
- Usually focuses on characteristics of:
  - People (social normative, social disorganization theories)
  - Places (routine activities theory)
  - Interactions of people and places (crime potential, ecology of crime)
- Alcohol distribution of interest since it is regulated and we have data on what and how much is sold where.

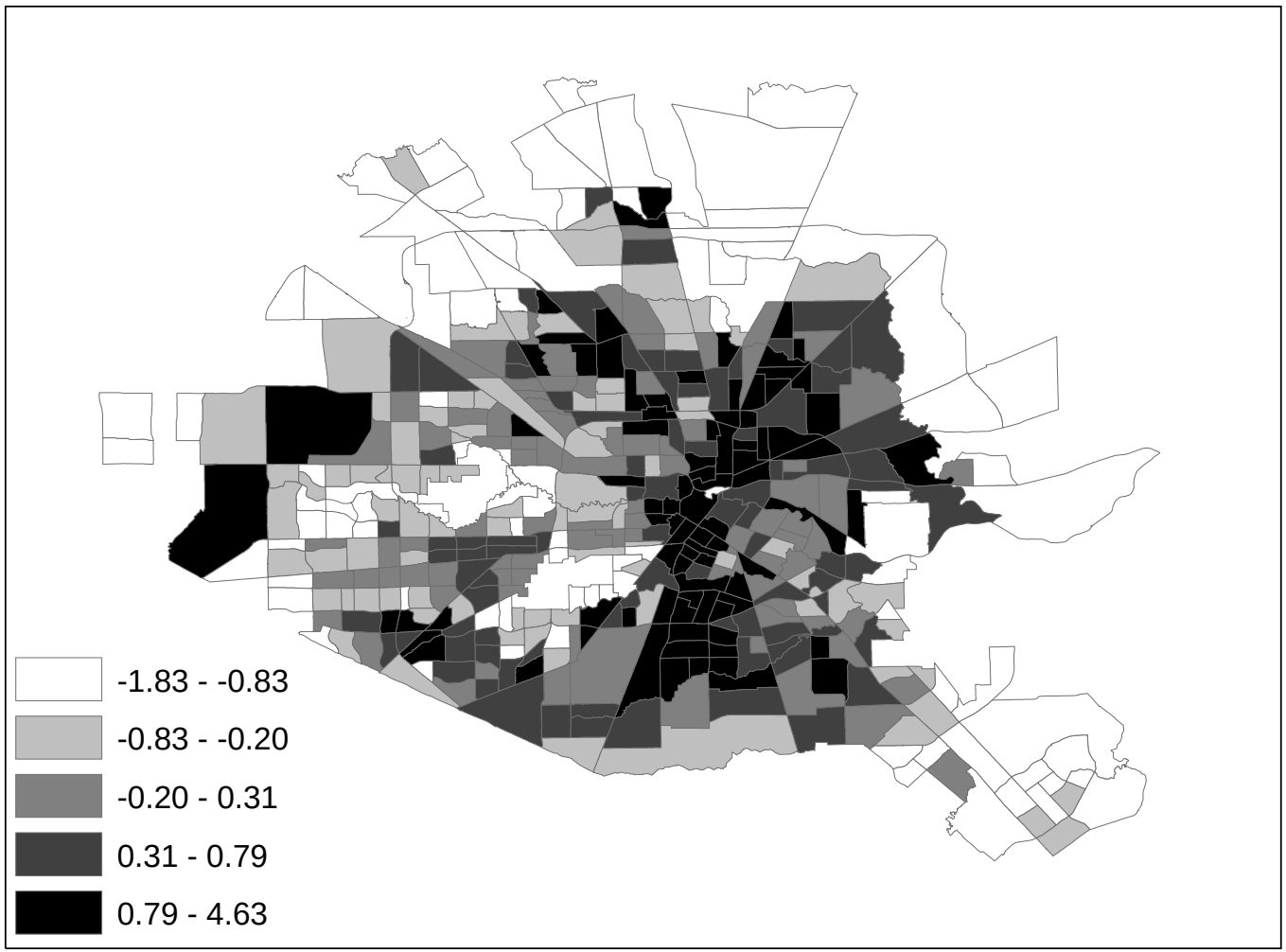
Data description

- Spatial support: 439 census tracts (2000 Census).
- Violent crime (murder, robbery, rape, aggravated assault) "first reports" for year 2000 from City of Houston Police Department website.
- Gorman et al. (2005, *Drug Alcohol Rev*) report less than 5% discrepancy with 2000 Uniform Crime Reports.
- 98% of reports geocoded to the census tract level.
- Alcohol data (locations of active distribution sites in 2000) from Texas Alcoholic Beverage Commission (6,609 outlets), 99.5% geocoded to the tract level.
- Drug law violations (also from City of Houston police data). 98% geocoded to the tract level.

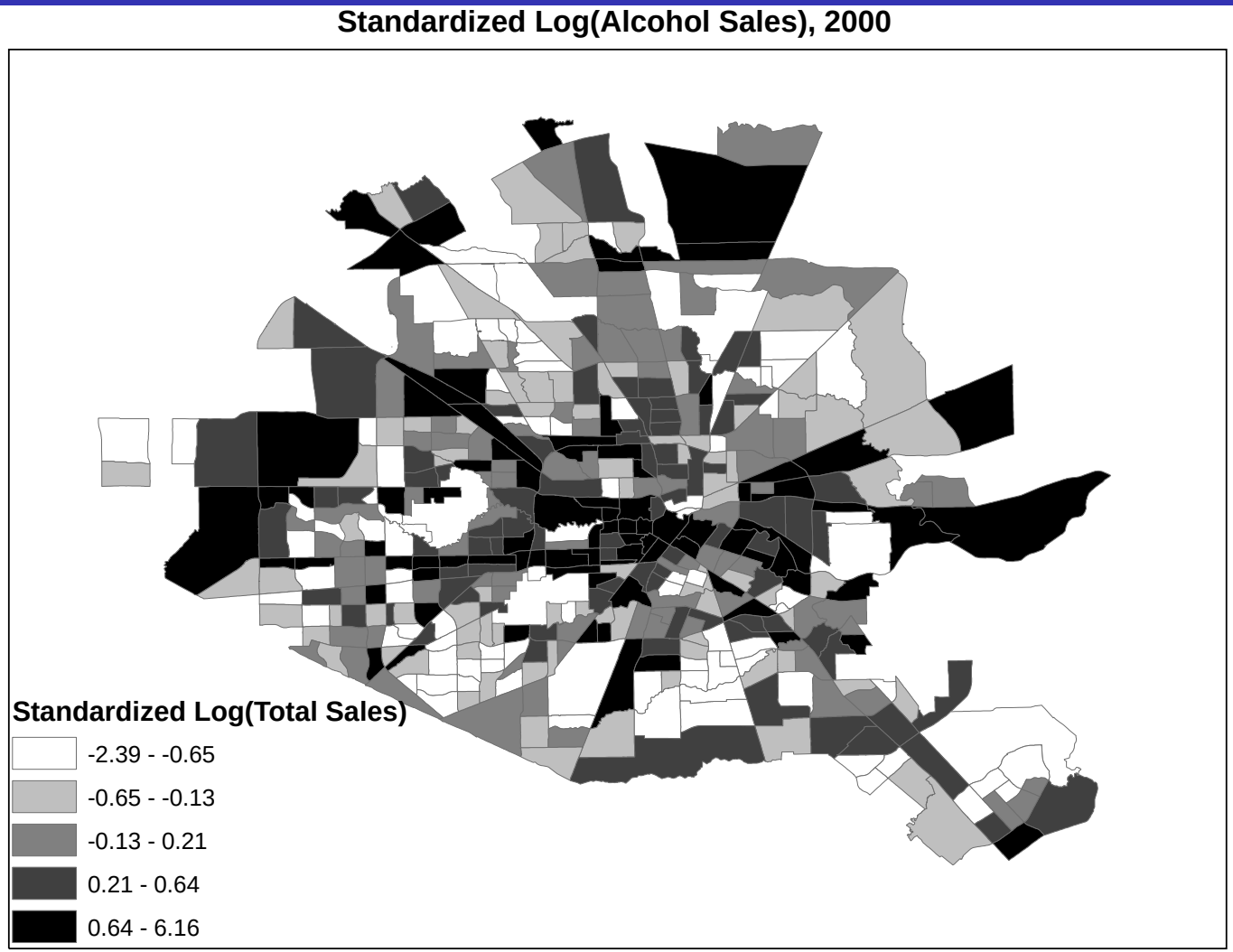
Violent Crime reporting rates, Houston, 2000



Standarized log(drug arrests), Houston, 2000



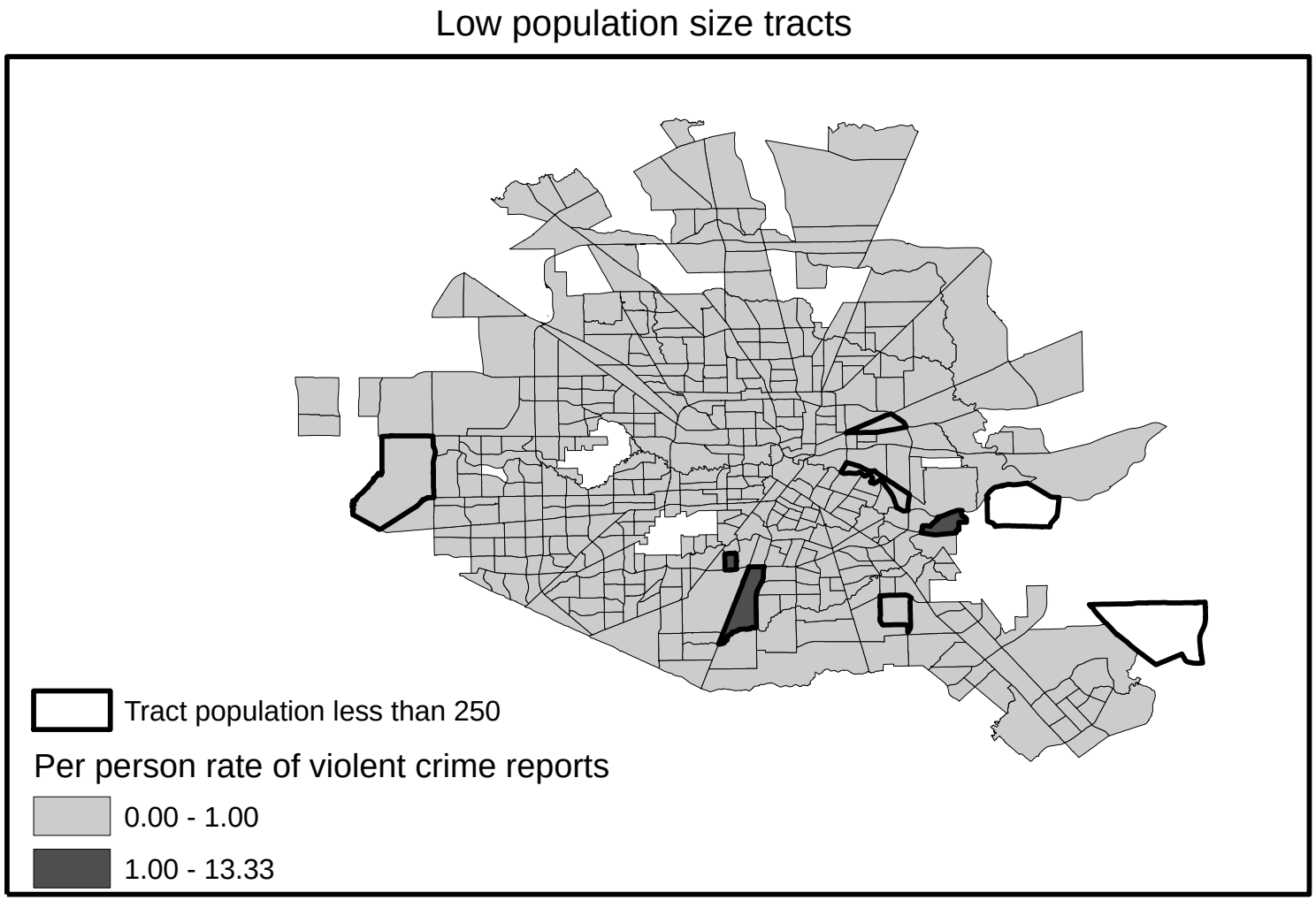
Standarized log(alcohol sales), Houston, 2000



Data "features"

- 7 of 439 tracts have extremely small population sizes: 1, 3, 4, 16, 34, 116, and 246.
- Tracts typically have 3,000-5,000 residents.
- Local rates for such tracts are extremely unstable (e.g., 40 reports, 3 residents).
- Actually a motivating a reason for including the spatially varying intercept: borrow information across regions.

Low population tracts and high rates



Basic Poisson regression

- Let Y_i = number of reports in tract $i, i = 1, \ldots, 439$.
- Suppose $Y_i \sim \text{Poisson}(E_i \exp(\mu_i))$, where E_i = the "expected" number of reports under some null model.
- Typically, $E_i = n_i R$ where all n_i individuals in region i are equally likely to report.
- $\exp(\mu_i)$ = "relative risk" of outcome in region i.
- We add covariates in linear format (within $\exp(\cdot)$): $\mu_i = \beta_0 + \beta_1 x_{alc,i} + \beta_2 x_{drug,i}$.
- Same "skeleton" for both GWR and SVC.

Why do we have E_i?

- $E_i = n_i R$ represents an "offset" in the model and lets us use Poisson regression to model *rates* as well as *counts*.

$$\begin{aligned} E[Y_i] &= E_i \exp(\beta_0 + \beta_1 x_{alc,i} + \beta_2 x_{drug,i}) \\ &= \exp(\ln(E_i) + \beta_0 + \beta_1 x_{alc,i} + \beta_2 x_{drug,i}) \\ &= \exp(\ln(n_i) + \ln(R) + \beta_0 + \beta_1 x_{alc,i} + \beta_2 x_{drug,i}) \\ \log(E[Y_i]) &= \ln(n_i) + \ln(R) + \beta_0 + \beta 1 x_{alc,i} + \beta_2 x_{drug,i} \end{aligned}$$

- GWR offset: $\ln(n_i)$, SVC offset: $\ln(n_i) + \ln(R)$.

GWPR (Nakaya et al., 2005)

- Geographically weighted Poisson regression.
- $\widehat{\beta}_{GWPR} = (\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{s})\boldsymbol{A}(\boldsymbol{s})\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{s})\boldsymbol{A}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})$.
- $\boldsymbol{A}(\boldsymbol{s})$ = diagonal matrix of Fisher scores.
- $\boldsymbol{Z}(\boldsymbol{s})$ = Taylor-series approximation to transformed outcomes.
- Update $\boldsymbol{A}(\boldsymbol{s})$, $\boldsymbol{Z}(\boldsymbol{s})$ and $\widehat{\beta}_{GWPR}$ until convergence.

Fitting in *R*

- Waller et al. (2007) use `GWR 3.0` software.
- In `R`: `maptools` will read in `ArcGIS`-formatted shapefile (files) into *R*.
- `spgwr` fits linear GWR and GLM-type GWR.

SVC

- $\mu_i = \beta_0 + \beta_1 x_{alc,i} + \beta_2 x_{drug,i} + b_{1,i} x_{alc,i} + b_{2,i} x_{drug,i} + \phi_i + \theta_i$.
- $\beta_0, \beta_1, \beta_2 \sim$ Uniform.
- Random intercept has 2 components (Besag et al. 1991):

$$\theta_i \overset{ind}{\sim} N(0, \tau^2)$$

$$\phi_i | \phi_j \sim N\left(\frac{\sum_j w_{ij} \phi_j}{\sum_j w_{ij}}, \frac{1}{\lambda \sum_j w_{ij}} \right).$$

  where w_{ij} defines neighbors, and λ controls spatial similarity.
- θ_i allows overdispersion (smoothing to global mean).
- ϕ_i follows conditionally autoregressive distribution (smoothing to local mean), generates MVN but more convenient for MCMC.

Defining the SVCs

- $\boldsymbol{b}_1, \boldsymbol{b}_2$ also given spatial priors and allowed to be correlated with one another.
- We use a formulation by Leyland et al. (2000) which defines

$$(b_{1,i}, b_{2,i})' \sim MVN((0,0)', \boldsymbol{\Sigma})$$

.

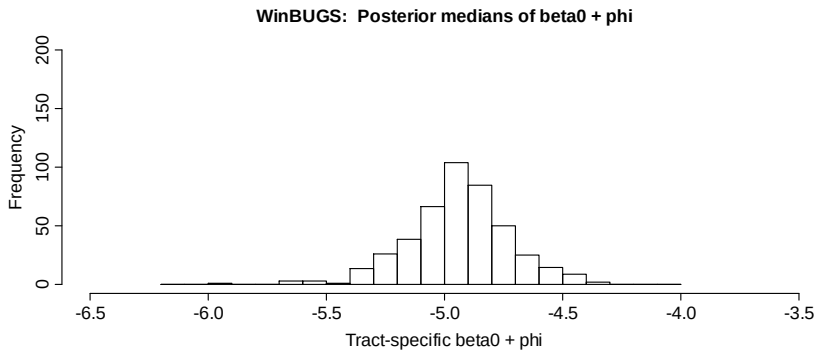
Fitting in *WinBUGS*

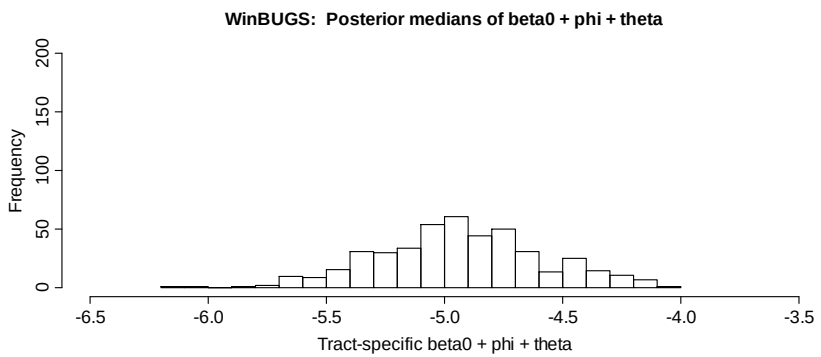
- Waller et al. (2007): Define the model in WinBUGS.
- MCMC fit.
- Note: Runs slooooooooowly.

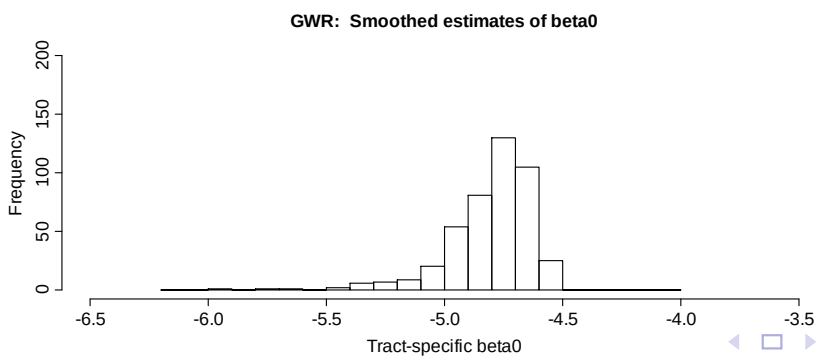
Implementation

- Waller et al. (2007): `GWR3.0` used to fit the GWR Poisson model.
  - Converged to estimate in $\sim$ 100 iterations.
  - Minutes.
  - Example code using `spgwr` library in `R`.
- `WinBUGS 1.4.1` used to fit SVC model.
  - Converged to distribution in $\sim$ 2,000 iterations.
  - 8,000 additional iterations used for inference.
  - Hours.
- Fit several versions of SVC model and compared fit via deviance information criterion (Spiegelhalter et al., 2003).
- Best fit included spatial varying coefficients, random intercept, and correlation between alcohol and drug effects.

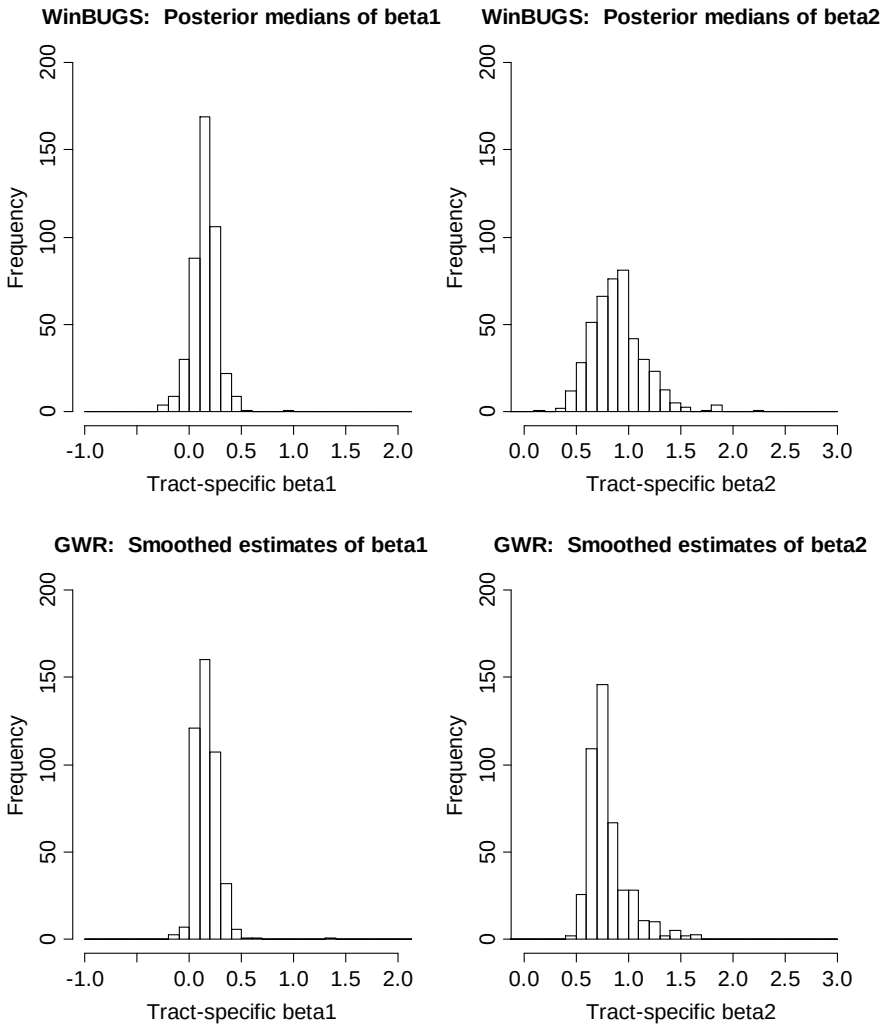
Results: Intercept





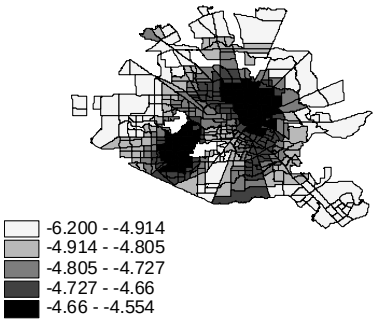


Results: Alcohol sales and drug arrests

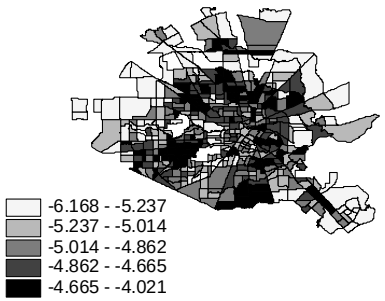


Estimated effects

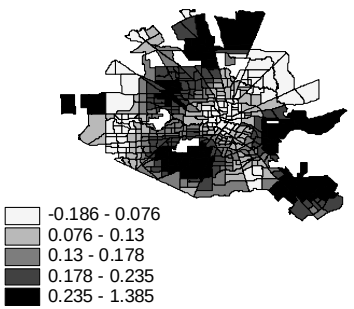
GWR: Local Estimate of Intercept



WinBUGS: Local Estimate of Intercept

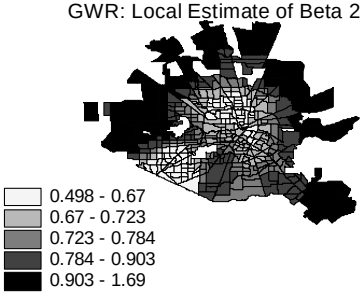


GWR: Local Estimate of Beta 1

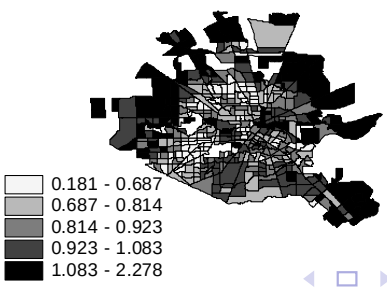


WinBUGS: Local Estimate of Beta 1

GWR: Local Estimate of Beta 2



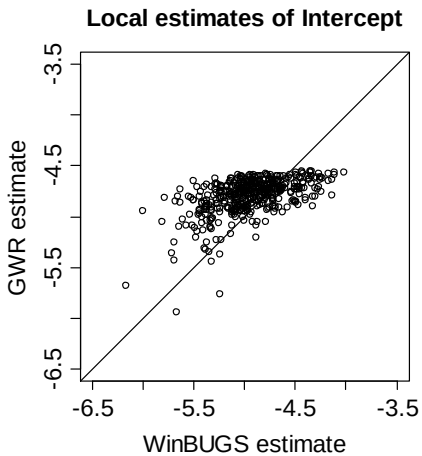
WinBUGS: Local Estimate of Beta 2

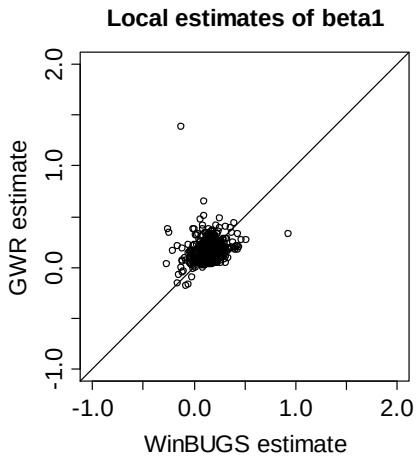


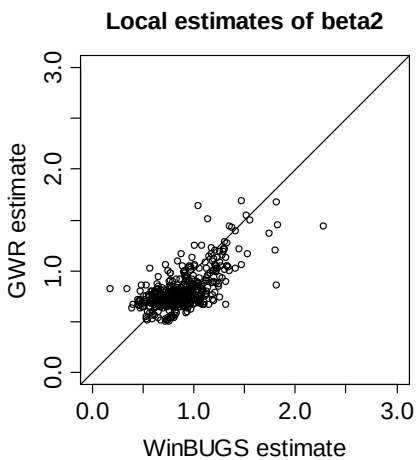
Similarities

- Alcohol: Increased impact in western, south-central, and southeastern parts of city.
- Illegal drug: Increased impact on periphery, lower influence in central and southwestern parts of city.
- Intercept: Increased risk of violence in central area, above and beyond that predicted by alcohol sales and illegal drug arrests.
- But, associations not too close...

Results: tract-by-tract



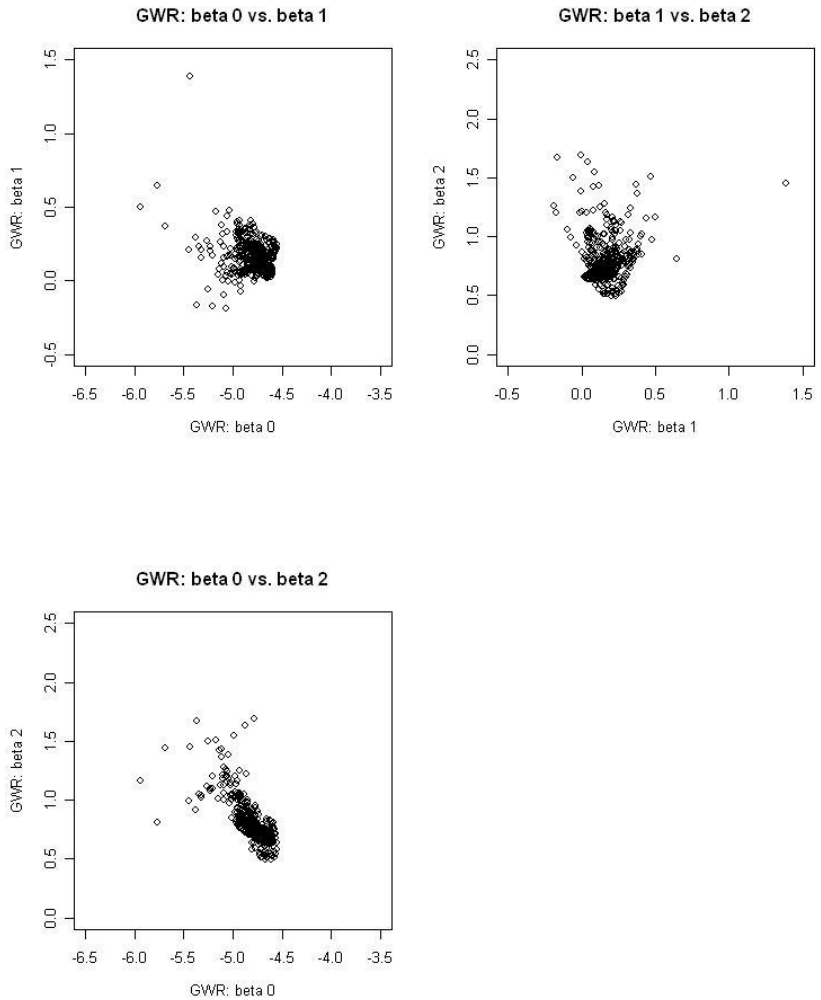




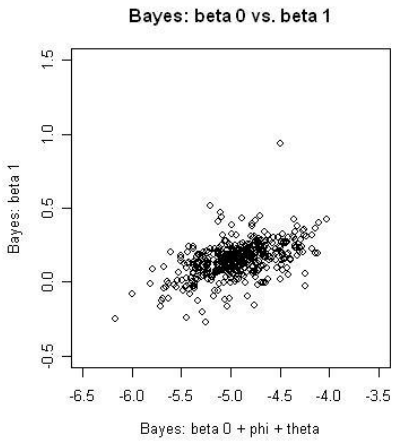
Differences

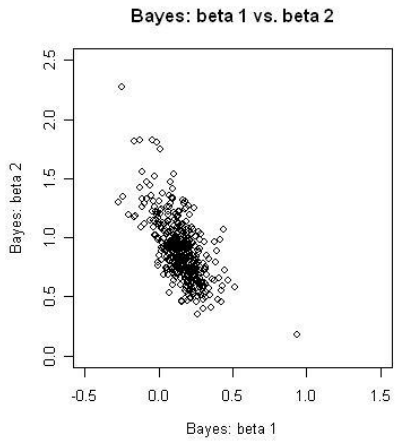
- GWR much smoother based on global best fit for *bw*.
- SVC used adjacency-based smoothing and a different amount of smoothing for each covariate.
- GWR: collineary between surfaces (Wheeler and Tiefelsdorf, 2005).
- SVC: Model based approach removes (or at least reduces) collinearity.

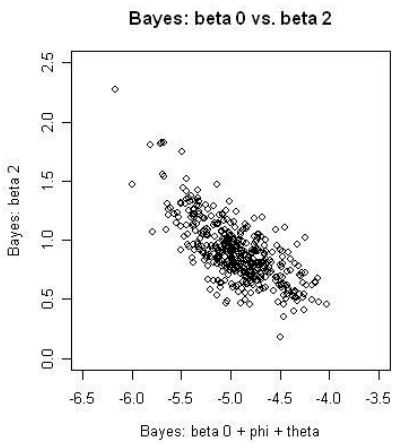
GWR: Collinearity?



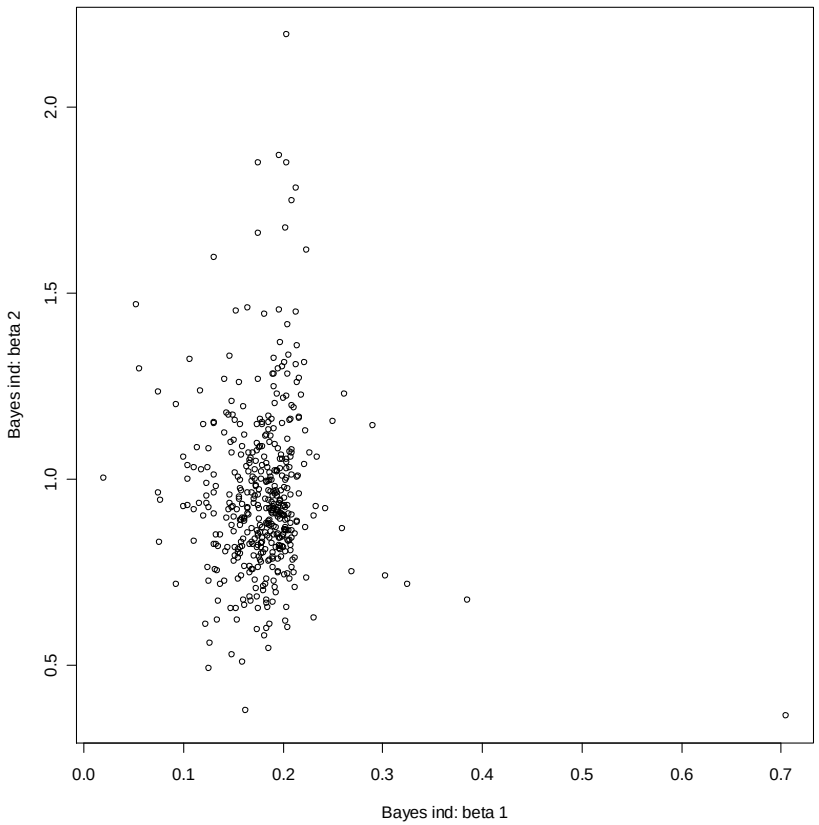
SVC: Collinearity?







SVC: No prior correlation



Let's try it out!

- Houston data on violent crime, alcohol sales, and illegal drug arrests.
- `ArcGIS` shapefile.
- Required `R` libraries: `maptools` (to read in shape file), `RColorBrewer` (to set colors), `classInt` (to set intervals of values for mapping), and `spgwr` (for GWR).

Conclusions

- GWR and SVC very different approaches to the same problem.
- Qualitatively similar in results, but not directly transformable.
- GWR fixed problems within somewhat of a black box.
- SVC allows probability model-based inference with lots of flexibility but at a computational cost (both in set-up and implementation).
- Further research:
  - Wheeler and Waller (2009): Attempt to set up SVC model to more closely mirror amount of smoothing in GWR.
  - Collinearity "ribbons".
  - Griffith (2002) eigenvector spatial filtering to adjust collinearity. Interpretability?