# Copula Modeling for Clinical Trials

Nathan T. James

October 2, 2018

## 1 Introduction

### Subsection

A significant body of research exploring the relationship between the environment and physical activity has developed in diverse fields such as urban planning, economics, criminology, transportation, psychology, exercise science, and public health. Research ranges from macro-scale studies of transportation mode choices across cities and regions to small scale investigations of specific types of walking behavior among small subpopulations. In public health, efforts have been made to determine correlates of purposeful physical activity such as walking or bicycling to school or work and the environment, since active transport is seen as a potential key to increasing physical activity and lowering rates of obesity and related chronic diseases. This is especially true for youth, for whom rates of physical activity have continued to fall as obesity rises. Additional research has focused on the relationship between active transport and overall levels of physical activity.

### Another subsection

Bla bla bla this is an example of a citation [5]

We can also include a picture as seen in figure 1. it's a placeholder for now

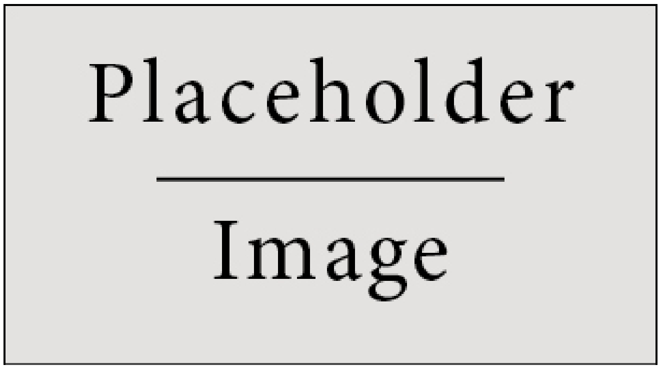

Figure 1: Some Figure

Table 1: NIfETy Incivility Items

| | |
|---|---|
| Total Broken Windows | People Swearing |
| Eviction Signs | People in Physical Fights |
| Un-Boarded Abandoned Buildings | People Loitering |
| Boarded Abandoned Buildings | Intoxicated People |
| Structures with Broken Windows | Evidence of Prostitution |
| Memorials | |

### Last Subsection

Multiple Opportunities to Reach Excellence (MORE) is a longitudinal, epidemiological study designed to assess the relationship between youth and long term exposure violence. The study participants are children in 4th and 5th grades in six Baltimore City public schools. In-person interviews were conducted with participating children and their parents to collect data on demographics, child activity (including child and parent reported walking to school behavior), and perceptions of the surrounding neighborhood.

$$q_j = \frac{\sum_i b_i p_i}{\sum_i l_i p_i} \tag{1}$$

Data was collected from 365 children in August through October 2007 as part of cohort 1 of the MORE study, however only 362 of the study participants had addresses which were successfully geocoded [6].

## 2 Next Section

### GIS methods

Here is some additional text and we can also reference a table (see table 1). The incivility was the sum of features present within a block [3].

The geocoded Nifety data was read into R and used to develop a spatial prediction model (described in next section). The predicted values and variances from these models were imported back into ArcGIS as a raster dataset The centroid of each street segment was calculated and the raster with predicted incivilities was joined with the centroids ([2]). Using a unique identifier, the centroids were rejoined to the street segments. For blocks on which NIfETy data was collected, the block quality was the actual observed incivility value. For unsampled blocks the block quality was defined as the predicted value at the street centroid.

Next, the MORE data for the children and schools was geocoded. 340 of the children's home addresses were geocoded automatically and another 22 were manually geocoded. Two of the remaining records had no information on the school and were removed. The address could not be geocoded for the last remaining record.

Finally, the MORE data with the attached path qualities was joined to the NSA data so that neighborhood level effects could be included in the final models. This complete dataset was exported from ArcGIS to R for analysis.

### Predicting Incivilities

The incivilities score for missing blocks was predicted using an ordinary kriging model. Ordinary kriging is an method for optimal prediction of spatially continuous processes which assumes a constant, unknown mean. For our case we consider points $\mathbf{z} = (z(s_1), \ldots, z(s_n))$ where $n$ is the total number of sampled spatial locations and the $s_i$, $i = 1, \ldots, n$ are an index of the spatial locations for the sampled points. Each $z(s)$ is an observed value of $\{Z(s) : s \in D\}$, which is a random process that varies continuously throughout the region $D \subset \mathbb{R}^2$ according to the spatial index, $s$

As in regular regression, we want to estimate 'large scale'/trend effects and small scale 'variance/covariance' effects. With just one observation from a joint probability dist and no assumptions, inference is not possible, since we need to estimate a potentially different mean for each location in addition to different covariance between any two locations. Therefore, we formulate a model and specify several assumptions which allow us estimate the model parameters. A process is (second order) stationary if the conditions below are met:

$$E(Z(s)) = \mu(s) = \mu, \text{for all } s \in D \tag{2}$$

$$Var(Z(s)) = \sigma^2(s) = \sigma^2, \text{for all } s \in D \tag{3}$$

$$Cov(Z(s_1), Z(s_2)) = C(s_i - s_j), \text{for all } s_i \neq s_j \in D \tag{4}$$

Equations 2 and 3 ensure that the mean and variance are constant and independent of location throughout the region $D$. Equation 4 ensures that the covariance depends only on the difference between two locations, rather than the locations themselves. $C(\cdot)$ in equation 4 is called the covariance function.

Another possible assumption related to the variability between points is

$$Var(Z(s_1) - Z(s_2)) = 2\gamma(s_i - s_j), \text{for all } s_i \neq s_j \in D \tag{5}$$

$2\gamma(\cdot)$ is called the variogram function and $\gamma(\cdot)$ is the semivariogram.

A more general form of stationarity called intrinsic stationary is equivalent to accepting the assumptions in equations 2,3 and 5. The form of stationarity chosen for model will determine which of equation 4 or 5 is used to model spatial dependence between points. Since the intrinsic stationarity assumption is more general and because the variogram and semivariogram have been shown to be more robust than the covariogram (citation) we will work with the semivariogram.

A further assumption called isotropy replaces the difference vectors $(s_i - s_j)$ with $||s_i - s_j|| = \mathbf{h}$ and assumes that that covariogram or variogram are functions of distance alone rather than distance and direction.

Plotting the values of the semivariogram against the distances for all pairs of locations, $\mathbf{h}$ gives a plot called the variogram cloud. (empirical variogram)

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(s_i) - Z(s_j))^2 \tag{6}$$

Where $\mathbf{h}$ indexes each class ($\mathbf{h}$ is often the midpoint of each class range), $N(\mathbf{h})$ is the set of all points within the class, $|N(\mathbf{h})|$ is the number of points within the class. By collecting the points within classes and then taking the average within each class we produce the empirical semivariogram estimator.

The empirical semivariogram is useful to explore the spatial dependence in the data, however in order to estimate values with the ordinary kriging equation, we must have a model for the 'true' variogram. Several families of variogram model have been validated and are considered ([7], [1]). To fit the model we rely on least squares methods which fit the model to the empirical semivariogram.

We begin the kriging process by importing the geocoded NIfETy data into R. We use the R package “copula” [4]. A map of NIfETy sample locations and incivility values was produced to visually examine spatial dependence. Next the variogram cloud, classical empirical semivariogram estimator, and robust semivariogram estimator were plotted and compared. Anisotropy was explored using directional semivariograms and an empirical semivariogram contour map.

# References

[1] Swati Biswas, Diane D Liu, J Jack Lee, and Donald A Berry. Bayesian clinical trials at the university of texas m. d. anderson cancer center. 6(3):205–216.

[2] Maria J. Costa and Thomas Drury. Bayesian joint modelling of benefit and risk in drug development.

[3] Maria J. Costa, Weili He, Yannis Jemiai, Yueqin Zhao, and Carl Di Casoli. The case for a bayesian approach to benefit-risk assessment:: Overview and future directions. 51(5):568–574.

[4] Marius Hofert, Ivan Kojadinovic, Martin Maechler, and Jun Yan. copula: Multivariate dependence with copulas.

[5] Harry Joe. *Dependence modeling with copulas*. Number 134 in Monographs on statistics and applied probability. CRC Press, Taylor & Francis Group.

[6] Roger B. Nelsen. *An introduction to copulas*. Springer series in statistics. Springer, 2nd ed edition.

[7] Michael Stanley Smith. Bayesian approaches to copula modelling. page 33.

# Appendix

## Appendix Subsection 1

Next, a streetmap was developed based on the 2007 TIGER/Line shapefiles. The TIGER shapefiles containing the location of streets, rail lines, and other passages in Baltimore city were merged with an accompanying relationship table containing information about street name, direction, zip code, and address ranges for each block face as well as indicator fields for type of linear feature (road, rail, other). Utilizing these indicator fields, the dataset was narrowed to include only road segments.

## Appendix Subsection 2

something more