

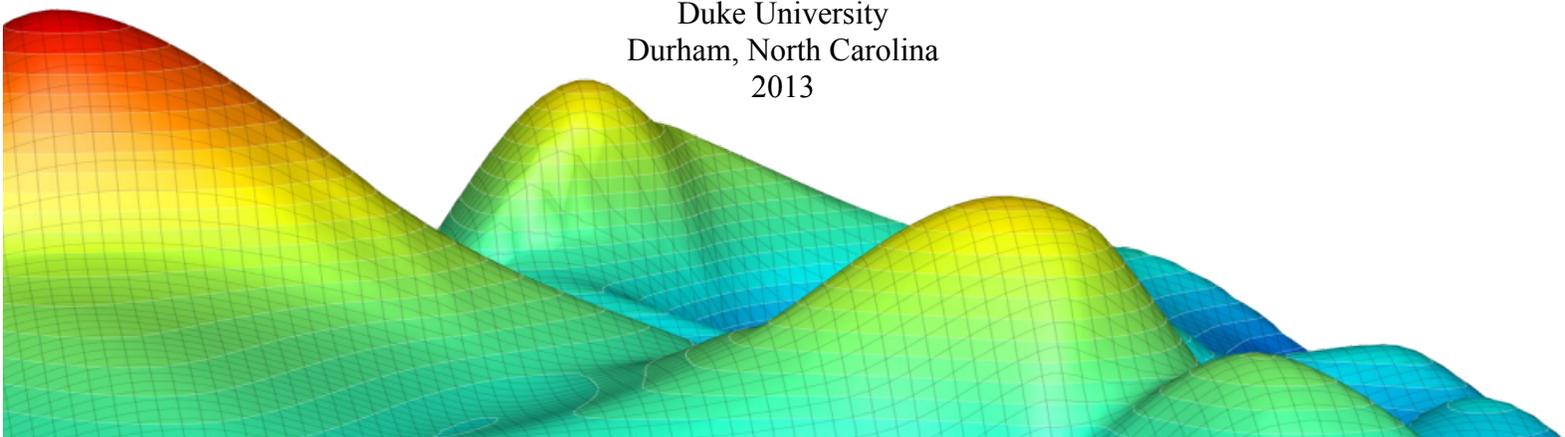
Geo-Spatial Modeling of Online Ad Distributions

Mitchel Drake Gorecki

Dr. Charles Becker, Faculty Advisor

*Honors thesis submitted in partial fulfillment of the requirements for Graduation with
Distinction in Economics in Trinity College of Duke University.*

Duke University
Durham, North Carolina
2013



Acknowledgements

Special thanks to Dr. Charles Becker for advising this work, Michael Els for providing the opportunity to undertake this project and MaxPoint Interactive for supplying the platform and support needed to make this research possible.

Abstract

The purpose of this document is to demonstrate how spatial models can be integrated into purchasing decisions for real-time bidding on advertising exchanges to improve ad selection and performance. Historical data makes it very apparent that some neighborhoods are much more interested in some ads than others. Similarly, some neighborhoods are also much more interested in some online domains than others, meaning viewing habits across domains are not equal. Basic data analysis shows that neighborhoods behave in predictable ways that can be exploited using observed performance information. This paper demonstrates how it is possible to use spatially correlated information to better optimize advertising resources.

JEL classification: C3; C33; C53; M37

Keywords: Spatial; Advertising; Online; Ad Distribution; Real Time Bidding

Table of Contents

1. Introduction	5
2. Spatial Modeling	6
2.1 Linear Regression Model	6
2.2 Spatial Model	7
2.3 Spatial Autoregressive Model (SAR)	10
2.4 Spatial Lag of X_i Model (SLX)	11
2.5 Time Lag Model	12
3. Measures of Spatial Autocorrelation	13
3.1 Moran's I	14
3.2 Geary's C	14
4. Theoretical Example	15
5. Direct Application	19
6. Data Source and Implementation	19
7. Backtest Results	20
8. Live Test Results	22
9. Known Issues/Comments	26
9.1 Measurement Metric	26
9.2 Low Spatial Correlation	27
10. Discussion of Results	27
11. Conclusion	30
12. References	32

1 Introduction

The purpose of this document is to explore the markets of real-time bidding (RTB) exchanges of online advertising. These exchanges are fast paced auctions for ad space on various websites. Unlike standard stock markets, these single advertisements are used only once for a single ad and cannot be resold. Related to online content buying is the fact these ads are spread across the country. It is possible to evaluate the performance of these ads by comparing their relative geo-spatial distributions. Since the country has a finite number of inhabitants, a continuous analysis of a surface across the country is not reasonable. Similarly, operating on a per person scale is too fine of a resolution. Such a scale is computationally expensive and tends to wash out any observable spatial trends. Instead, the country can be divided into 52,000 neighborhoods that represent groups of people spatially related to each throughout the day. These neighborhoods are determined by grouping individuals that share similar traits such as socioeconomic status, race or any other factor. This number of divisions creates an optimal coarseness of regions that people tend to group. If too many divisions are used, the model will not accurately predict spatial elements. Having too many groups introduces too much noise into the data making it difficult to identify trends in data. Using too few divisions creates an almost uniform surface purged of any usable trends. 52,000 neighborhoods is a fine enough scale that separates dissimilar neighbors into subsets without adding too much noise. As geo-correlated data becomes more readily available, additional divisions will become increasingly more useful. When individuals move in and out of these neighborhoods, people naturally tend to segregate themselves spatially and mentally. Having record of the segregation allows for the optimization of targeted advertising distribution based on location.

Essentially, the concept is to use historical data to predict the optimal manner in which to successfully distribute future ad space. The only requirement for a success metric is that it is quantitative. The analysis was performed in conjunction with MaxPoint Interactive, a real time bidding advertising company. MaxPoint provided the means by which to instantly adjust the purchasing and ad distribution algorithm to observe how spatial effects can be used to optimize ad campaigns. Tests performed in this document use the standard metrics of click through rate, view through rate and pixels as examples. Clicks represent when an ad is clicked on and linked to a new page. A view is when the consumer interacts with the ad in a manner that indicates they

have acknowledged the ad. A pixel is a term for any other metric that might be used, such as if the consumer hovers a mouse over the ad or triggers a video ad to play. Through initial analysis of the data, it was observed that a view through rate is a better metric as it is observed more frequently. This means that view through rates had the least noise or misleading data. The optimization of ad space allocation is done by identifying what attributes of a consumer and ad produce the most successful outcome. Currently, the method is to produce a regression of predicted outcomes that can be used to appropriately allocate advertising resources. This regression uses independent variables such as consumer income levels, housing values, or educational zoning. Locations with low predicted success rates are not advertised to and those with higher success are focused on. Currently, the model does not consider spatial correlation, a potentially vital component of the residuals of this regression. Spatial correlations between the ad and consumer preferences can be used to reduce the error associated with the current model. Considering this idea leads to an optimal geo-spatial ad distribution that both lowers cost by reducing less successful ads while improving success rates by catering ads to areas in which they are most effective. Thus, the final goal is to improve the current model to incorporate spatial techniques to further explain the residual error. The next portion of this document will review various spatial models and their strengths and weaknesses.

2 Spatial Modeling

This portion of the document is devoted to the introduction of spatial modeling. The purpose is to create a general understanding of the mechanics of spatial modeling and how it can be applied.

2.1 Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_1 X_1 + \dots \beta_n X_n + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2) \quad i = 1, \dots, n$$

This is a simple demonstration of a linear regression of n explanatory variables on a dependent variable Y_i . The vector X_n represents a set of explanatory variables with associated parameters, β_n . These explanatory variables consist of any relevant data correlated with ad

performance. The vector Y_i represents the performance of observed ads. Each observation has an underlying mean of $\beta_n X_n$ with a random component, ε_i . This error term, ε_i , is assumed to have the standard traits of an error term. This means the error is normally distributed with an average value of zero, and is completely uncorrelated to X_n or Y_i . Each observation i represents a set of points in space with observed values at those locations. This model can be modified to include more complex regressions such as logarithmic or higher order powers or moments of explanatory variables. This model does not take into account any spatial correlation between points and is assumed to be a basic representation of the current model under which an advertising agency may operate. Normally it is assumed that these points are statistically independent, implying that $E(\varepsilon_i \varepsilon_j) = E(\varepsilon_i)E(\varepsilon_j) = 0$. This is not true in a spatial context. Rather, the value of Y_1 depends on Y_2 and visa versa. For example, predicting if it will rain in your yard is a function of indicators like air pressure, air saturation, temperature, but most usefully knowing whether or not it is raining in your neighbor's yard. The spatial element can be used to better predict data beyond this simple regression model.

2.2 Spatial Model

By performing linear regressions between each point to the rest of the set, it is possible to build a system of equations that can evaluate spatial components of a data set. This is of little use though since it results in a system with more parameters than observations. Such a process is also time consuming and computationally intense, with n^2 relations to be considered. The solution to over-parameterization is to induce structure based on spatial relations. This is done by introducing parameters that are geospatially defined. This is the spatial autoregressive process.

$$Y_i = \rho \sum_{j=1}^n W_{ij} Y_j + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2) \quad i = 1, \dots, n$$

The particular form of this equation used can be found in LeSage and Pace (2009, p. 8). The idea is that Y_i is best predicted by some initial set of Y_j given spatial relations, W_{ij} . The intercept term is eliminated under the assumption that the vector of observations on the

independent variable are deviations from the mean of the entire set. The $\sum_{i=1}^n W_{ij}Y_j$ term is called a spatial lag since it represents a linear combination of Y_j values constructed from neighboring observations. The elements in W_{ij} are an n by n spatial weight matrix that determine how points in Y_j are related. So for instance, to predict the likelihood of someone owning a BMW, one can observe the cars their neighbors' drive and use the correlation between these neighborhoods to make a reasonable guess. The matrix G as seen below represents an example of an evaluated area. The area is shown by Y_{ij} values that represent some metric at a location. Each Y_{ij} value in this matrix is a success rate for a campaign at that location. For calculations, the matrix must be linearized into Y to be multiplied by W_{ij} .

$$G = \begin{bmatrix} Y_{11} & Y_{12} & Y_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ Y_{31} & Y_{32} & Y_{33} \end{bmatrix} \rightarrow Y = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{31} \\ Y_{32} \\ Y_{33} \end{bmatrix}$$

$$W_{ij}Y = \begin{bmatrix} 0 & 1 & 0 & 1 & .5 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & .5 & 1 & .5 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & .5 & 1 & 0 & 0 & 0 \\ 1 & .5 & 0 & 0 & 1 & 0 & 1 & .5 & 0 \\ .5 & 1 & .5 & 1 & 0 & 1 & .5 & 1 & .5 \\ 0 & .5 & 1 & 0 & 1 & 0 & 0 & .5 & 1 \\ 0 & 0 & 0 & 1 & .5 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & .5 & 1 & .5 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & .5 & 1 & 0 & 1 & 0 \end{bmatrix} * \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{31} \\ Y_{32} \\ Y_{33} \end{bmatrix}$$

The spatial model alone introduces the parameters W_{ij} and ρ , which respectively add a spatial component and weight indicating spatial importance. W_{ij} is an n by n matrix that contains the spatial correlation between each location in the model. Spatial analysis calls for the application of assumed spatial relations rather than measuring the correlation between each set of points as in a standard regression. This is because the purpose of a spatial analysis is to identify the spatial elements that determine an outcome to contribute to a regression analysis. The

weighting matrix can be specified in many ways, of which the examples shown here are from Brusilovskiy (2009):

- The weight for any two different locations is a constant
- All observations within a specified distance have a fixed weight
- K nearest neighbors have a fixed weight and all others are zero
- Weight is proportional to inverse distance, inverse distance squared, or inverse distance up to a specified distance
- Some other method

An explanation of this matrix is summarized here from LeSage and Pace (2009, p. 21). It is standard for a weighting matrix to have identical values mirrored across the diagonal. This makes sense since the relation from point A to B should be the same for point B to A. This means that when establishing the weighting matrix, one corner of the matrix can be left as zeroes, pulling data from the other corner of the matrix. This trick will increase computation efficiency but will not be used for discussion purposes. For simplicity, the example above shows a world where there is perfect correlation (1), marginal correlation (.5), or no correlation (0). Each Y_i contains the data for latitude, longitude and a performance value. Those points directly next to a given point are allotted perfect correlation, diagonals with .5 and with all others as 0. Every value shows how each Y_i impacts those next to it. So for instance, $W_{ij}Y_{13}$ is a function of $\rho(Y_{12} + .5[Y_{22}] + Y_{23})$. This calculated result can easily be seen by looking at matrix G, which maintains the visual spatial relations. The diagonal of matrix W is zeros since there is no spatial impact for a location on itself. The parameter ρ is used as a weighting, or correlation factor, to alter the degree of importance placed on the spatial component of the model.

The main drawback of such a model is that a large spatial correlation matrix with n^2 entries must be made where n is the number of locations. More importantly, such a model requires an initial value for each Y_i to spatially adjust. This is the equivalent of having a lagged dependent variable. The model requires the use of a dependent variable in the regression creating an endogenous problem. An example is where a current performance value for every point is needed prior to generating a spatially dependent prediction of those already known values. This

is a problem when attempting to predict the unknown performance of a location. Lastly, notice that the spatial arrangement of locations in G has been idealized as a square. This is almost never the case. For this reason, it is advised to either maintain a linearized location vector Y with a particular order or to construct a reference library for these values.

2.3 Spatial Autoregressive Model (SAR)

For this model, the parameter α is used to represent a vector of mean Y_i values for each point. This is multiplied by the identity matrix i_n to create a diagonal matrix of these values. The purpose of this is to accommodate for offset induced when the set Y does not have a zero mean value. The idea is to force the spatial matrix to evaluate changes in the data, rather than overall performance. The equation is seen below as it appears in LeSage and Pace (2009, p. 32).

$$\begin{aligned}
 Y_n &= \alpha i_n + \rho W_{ij} Y_n + \varepsilon_i \\
 (I_n - \rho W) Y_n &= \alpha i_n + \varepsilon \\
 Y_n &= (I_n - \rho W)^{-1} i_n \alpha + (I_n - \rho W)^{-1} \varepsilon \\
 \varepsilon_i &\sim N(0, \sigma^2 I_n)
 \end{aligned}$$

The expected value of each observation Y_i will depend on the mean value α plus a linear combination of the values of neighboring points scaled by ρ . This shows how the data are generated in a simultaneous, spatially autoregressing nature, where the model derives its name. An infinite series is generated by taking further powers of W as the model considers second, third and higher order neighbors. So the matrix W^2 reflects second-order contiguous neighbors. Since the second order neighbors are neighbors to the original i observation being considered, W^2 will have positive elements on its diagonal. If additional, higher orders of neighbors are considered, the solution generates an infinite series. When this series is evaluated to an infinite number of neighbors, the solution converges on the following equation.

$$Y_n = \frac{1}{1 - \rho} i_n \alpha + \varepsilon + \rho W \varepsilon + \rho^2 W^2 \varepsilon \dots$$

Although, considering an infinite number of neighbors would be impractical for the proposed problem. As such, it will later become an optimization problem to determine the best

order of neighbors to consider. The spatial autoregressive structure can be combined with a standard regression to produce the standard regression model shown here.

$$Y_n = \rho W_{ij} Y_n + \beta_i X_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

Although merely a synthesis of the two models above, this noteworthy expression is the foundation for useable spatial modeling functions. Originally, models that did not consider spatial factors produced biased estimates of β_i . The use of the spatial matrix W in combination with the linear regression alters the predicted parameters to better match their actual values independent of location. This is the type of model that is most applicable to the advertising market. As Gaetan and Guyon (2010) point out, the only problem is that the model still has the problem of endogeneity, containing a lagged variable. The parameters to be estimated are the standard regression parameters, β_i , ρ and σ . If the parameter ρ takes a value of zero, there is no spatial dependence, and the model is a standard regression.

2.4 Spatial Lag of X_i Model (SLX)

This model is similar to the SAR model except this time the independent variables are also spatially evaluated. Such a model would be appropriate for regressing housing values or even community value of pretty landscaping. The presence of pretty landscaping is positively correlated with housing values, but also impacts the value of the neighbor's homes. In this case both the value of the home as well as the presence of landscaping are spatially correlated to housing prices. The model is specifically written in LeSage and Pace (2009, p. 36) as:

$$Y_i = \alpha_i + \beta_i X_i + \gamma W_{ijz} X_i + \varepsilon_i$$

The initial linear regression is present with α representing a diagonal matrix containing intercept values. Instead of ρ , γ is now used since the correlation factor is now a vector with values for each X_i . W_{ijz} is now a three dimensional matrix with spatial weightings for each X_i . The biggest drawback is a significant increase in the computational power required to produce such a model. This is because each of the independent variables X_i is spatially evaluated between

themselves, requiring (n^2) correlations. This model is called the spatial lag of X model, or SLX, since the model contains spatial lags, WX, of neighboring characteristics as opposed to just observed outcomes. A benefit of using this model is that it provides a better understanding of how individual population characteristics determine the spatial distribution of ad performances. When only the performance metric is spatially regressed, as in the SAR model, the spatial component can be accounted for, but not explained. As discussed by Diggle, Fuentes, Gelfand and Guttorp (2010), the SLX model is able to indicate what parameters contribute to success in a spatial context. Knowing these relations can be of value, but it is significantly harder to implement an SLX model and is computationally taxing.

2.5 Time Lag Model

Most economic decisions are made by analyzing the behavior of relevant variables over a past time period. For instance, it is observed that children tend to most influence a parent's choice of cereal purchase when that desired cereal is in direct sight. Hence, sugary cereals marketed towards children tend to be at child-eye level, on lower shelves. The response of product placement is possible due to an analysis of the past sales observations. This time element can be mathematically modeled and introduced into a spatial regression.

Consider a relation where the dependent variable, Y_t , at time t is determined using a SAR model that depends on space-time lagged values of that variable from neighboring observations. This generates a time lag of the neighboring values of the dependent variable in the time period $t-1$. The result is a spatial weighting matrix $W_{ij,t-1}$. The final model as follows:

$$Y_{n,t} = \rho_t W_{ij,t-1} Y_{n,t-1} + \beta_i X_{i,t} + \varepsilon_{i,t}$$

This last generic model introduces the concept of time lagging. Historical data are used to create prediction parameters for future outcomes. This is especially useful when endogeneity is present in a model. The W_{t-1} matrix can be applied to the SAR model to predict future outcomes at time t . Since this matrix is created from historical data it must be assumed that the change in spatial properties between time intervals is not significant. Each subsequent time is predicted using the adjusted W matrix from the actual previous outcome. This model is essentially a time

lagged autoregressive model. It has the advantage of being able to regress into the future, predicting unknown outcomes from previous observations. This is the model on which the analysis will continue since it is indicative of the data being used.

Notice that a recursive substitution occurs when $Y_{n,t-1}$ is substituted with $\rho_{t-1}W_{ij,t-2}Y_{n,t-2} + \beta_i X_{i,t-1} + \varepsilon_{i,t-1}$. If this substitution is performed q times, the prediction of $Y_{n,t}$ becomes:

$$Y_{n,t} = (I_n + \rho W + \rho^2 W^2 \dots + \rho^{q-1} W^{q-1}) X \beta + \rho^q W_{ij,t-q}^q Y_{t-q} + u$$

$$u = \varepsilon_t + \rho W \varepsilon_{t-1} + \rho^2 W^2 \varepsilon_{t-2} \dots + \rho^{q-1} W^{q-1} \varepsilon_{t-(q-1)}$$

The expression can be simplified by taking the limit as q goes to infinity, yielding the result in LeSage and Pace (2009, p. 70):

$$\lim_{q \rightarrow \infty} E(Y_t) = (I_n - \rho W)^{-1} X \beta$$

This shows that a SAR model is possible from a time-dependent series of decisions for various points in space dependent on each points' neighbors.

3 Measures of Spatial Autocorrelation

The main problem with such models is determining if a spatial model is appropriate. Whenever additional parameters are added to a model, the correlation can only increase. This means that if a small, possibly random, improvement is observed, there will be false indications that a spatial model is useful. This means that a standard metric for determining spatial significance must be used.

3.1 Moran's I

Moran's I is a test for global spatial autocorrelation of a continuous data set. The test statistic is represented and explained below as shown in Lembo (2008, p. 10).

$$I = \frac{n \sum_i \sum_j (x_i - \bar{x})(x_j - \bar{x})}{S_0 \sum_i (x_i - \bar{x})^2}$$

S_0 is the sum of elements of the weight matrix W . Moran's I is similar to the correlation coefficient in that it varies over ± 1 . In the absence of autocorrelation, I has an expected value of $-1/(n-1)$, which converges on zero with large values of n . A value greater than zero indicates positive correlation and less than zero as negative correlation. The variation of Moran's I is:

$$\text{Var}(I) = \frac{n((n^2-3n+3)S_1-nS_2+3S_0^2)-k(n(n-1)S_1-2nS_2+6S_0^2)}{(n-1)(n-2)(n-3)S_0^2}$$

$$S_1 = \frac{1}{2} \sum_{i \neq j} (W_{ij} + W_{ji})^2 = 2S_0 \text{ for symmetric } W \text{ containing } 0\text{'s and } 1\text{'s}$$

$$S_2 = \sum_i (W_{i0} + W_{0i})^2 \text{ where } W_{i0} = \sum_j W_{ij} \text{ and } W_{0i} = \sum_j W_{ji}$$

In standard practices it is proper to evaluate the variation to ensure statistical significance is achieved. This prevents improper reporting of untrue results. The second method of evaluating spatial autocorrelation is Geary's C.

3.2 Geary's C

Geary's C is a test statistic based on the deviations in responses of each observation with another. The parameter is C such that:

$$C = \frac{n-1}{2S_0} \frac{\sum_i \sum_j w_{ij} (x_i - x_j)^2}{\sum_i (x_i - \bar{x})^2}$$

Geary's C ranges from 0 having maximum positive autocorrelation to 2 with max negative autocorrelation. A value of 1 is expected for a data set with no autocorrelation. If positive autocorrelation is observed, the variance of C is:

$$Var(C) = \frac{1}{n(n-s)(n-s)S_0^2} (S_0^2[(n^2 - 3) - k(n - 1)^2] + S_1(n - 1)[n^2 - 3n + 3 - k(n - 1)] + \dots$$

$$\frac{1}{4}S_2(n - 1)[k(n^2 - n + 2) - (n^2 + 3n - 6)])$$

S_0 , S_1 and S_2 are the same as for Moran's I. Geary's C is similar to Moran's I but is generally accepted as placing greater weight on local spatial correlation over global. Moran's I tends to be more sensitive to extreme values of Y_i where Geary's C is more sensitive to differences in small neighborhoods. This generalization is taken from Lembo (2008, p. 12), who explains in detail how, generally, Moran's I is preferred as it is consistently more powerful than Geary's C. As such, the metric of Moran's I will be relied upon as the more important spatial statistic.

4 Theoretical Example

This example will be a walkthrough of how spatial models can reduce the residuals of a regression. It is similar to the demonstration in Bivand, Pebesma and Gomez-Rubio (2007) except that theoretical data is used to make trends more obvious. Figure 1 below is an example of observed success rates over a given area. This could be any quantitative success metric, such as the initially observed click through rates (CTR) for a campaign over a given area outlined by the XY plane.

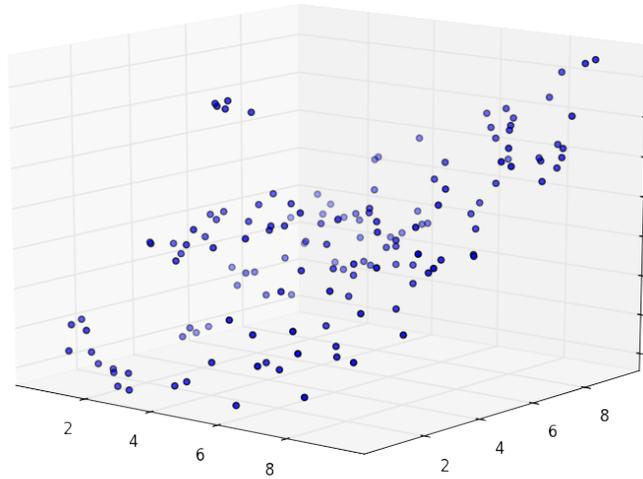


Figure 1: Initial observed data¹

The two main trends observed in the data are an upward sloping plane and a local maximum in the middle of the area. A firm's multivariable regression of this data might look similar to Figure 2 below. Spatial arrangement is currently not used in the prediction of campaign performance.

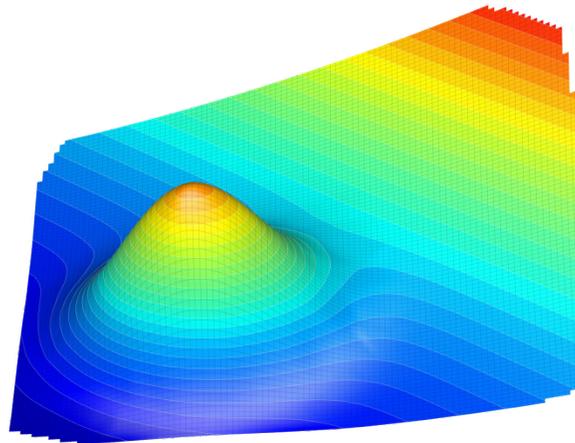


Figure 2: Initial regression

¹ Data modeled from the equation: $10 \cdot \exp(-5 \cdot (((x[i]-m/3)**2/(2 \cdot \text{std}(x)**2)) + ((y[i]-m/3)**2/(2 \cdot \text{std}(y)**2)))) + (x[i]+y[i])**2/25$ where x,y are random locations for given success rates

The local peak in this plot could possibly represent the location of a store. As the distance from the store increases, a customer's willingness to travel to make a purchase decreases. The overall increasing trend could be the transition into a neighborhood with different purchasing habits. The residuals of this regression can be seen in Figures 3 and 4. To the left is a similar 3D view and to the right is a downward view of the same residuals, but with the actual observed data points indicated on the plot. Using spatial correlation, it is possible to interpolate the value of points in between the actual observed data.

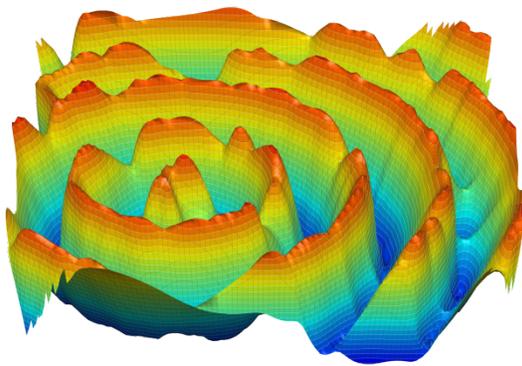


Figure 3: 3D residuals

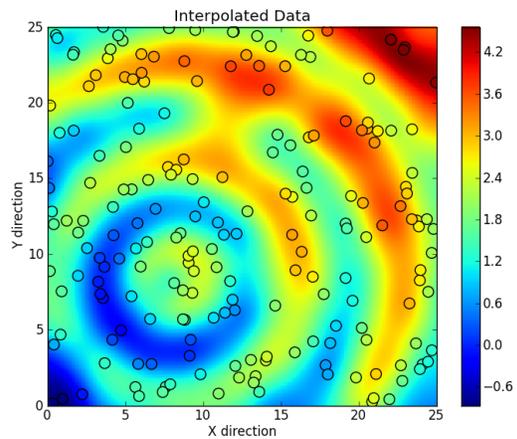


Figure 4: 2D residuals

The interpolation points out that there is a clear trend in the residuals. This is done to visually indicate the potential need for the construction of a spatial model from the original regression. What this indicates is that the original model is incorrectly estimating the click through rates for a given ad in a manner that could be explained spatially. Had a spatial element been considered, the model could have better predicted the CTR. Figure 5 shows the new model considering spatial relations.

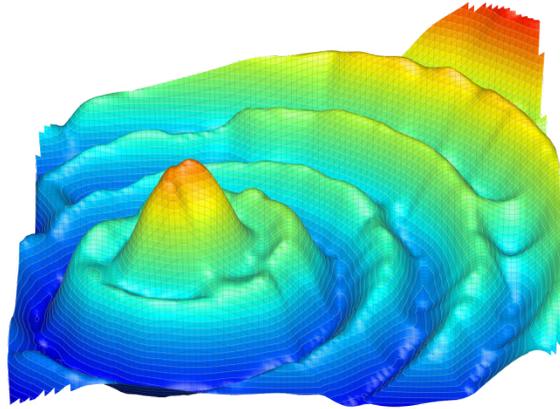


Figure 5: Spatial model

Notice how the spatial model appears to be a synthesis of the original model plus the weighted spatial elements in the old residuals. The new residuals between the spatially adjusted model and the observed data can be seen in Figure 6.

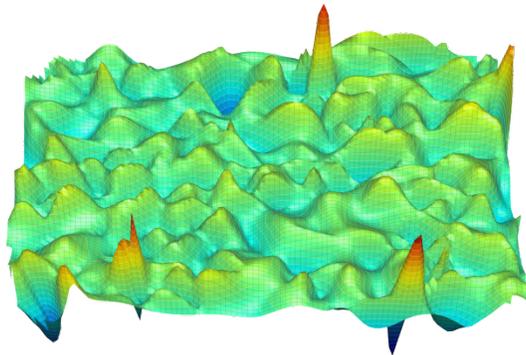


Figure 6: New residuals

The key point is that the new residuals are almost perfectly randomly arranged and less in magnitude than before. This indicates that there is significantly less unobserved trend in the data and that the spatial model was both appropriate and successful.

5 Direct Application

Spatial modeling cannot be used in the manner presented to solve the problem of optimally distributing ads. Rather than trying to best model a surface of success rates over an area, the objective is to be able to predict which locations would perform best given a partial set of success rates. This means that rather than interpolating the spatial prediction at every point on a surface, the computation is reduced to predicting values at digital zips not already campaigned in. This is because rather than a continuous surface, the country is divided into discrete locations called digital zips. The estimated values of these points are compared to determine which would be the best to add to a campaign. This ‘twist’ to the basic spatial modeling technique is demonstrated in a backtested example below.

6 Data Source and Implementation

The remainder of this document uses data provided by MaxPoint Interactive. Data was provided in the form of SQL tables and is manipulated using Python with the Pysal plugin. The work of Anselin (2005) served as a reference for gaining familiarity with the implementation of spatial algorithms. The relevant data included items such as locations, success rates, city divisions, school zones, income distributions, and other population characteristics organized by geography. Administrative access to ad campaign locations was granted to allow for the alteration of ad distribution. The analysis of spatial dependence began with observing success metrics over different regions of the US for various ads. The proprietary ‘digital zips’ (DZs), or neighborhoods, were mapped according to latitude, longitude and their success rates. The rates used were clicks, views, pixels or specific user actions on a webpage. All of the data for each individual ad were aggregated and regressed to determine the presence of spatial correlation. Python code was written to enter the SQL servers, analyze a campaign for spatial relations and to adjust the distribution of ads accordingly. Changes in performances of new ad locations were observed in real time, with useful results yielded by the end of each day. Further optimizations were performed to hone the algorithms to better seek out the best locations to advertise as well as those locations to avoid. These optimizations ranged from determining the best number of neighbors to consider, how large of a radius those neighbors must be within or how data

clustering should be handled. Specifics regarding how each of these elements impacted the algorithm are discussed in the optimization section. Before live campaigns were adjusted, backtesting was used to verify the functionality of the spatial model.

7 Backtest Results

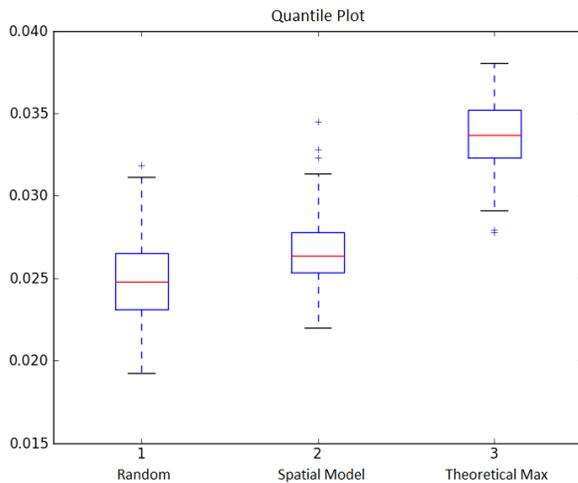
The purpose of backtesting is to show how spatial modeling could be used to optimize currently running campaigns, without accidentally harming the success of a running campaign. The goal is to use already observed data to predict the next optimal digital zip to advertise in. Ads are catered to the areas that have the highest likelihood of success and potentially stopped in those areas of low success. The results from adding and removing digital zips from a campaign are monitored for accuracy. Over time, this model learns and adapts to target areas of highest success rates.

Backtesting was performed by taking past data from an observed video campaign, randomly removing a percentage of the data and then attempting to spatially predict which points performed the best of the points removed. This backtest was performed on a real, past campaign for four different spatial weighting methods with the purpose of selecting the optimal weighting method. The methods tested were those found in Gaetan and Guyon (2010):

- Constant weighing for k nearest neighbors
- Constant weighting of neighbors within a fixed radius
- Discounted distance weighting for k nearest neighbors within a fixed radius
- Considering congruent neighbors with higher order neighbors having decaying weight
- A combination of these or other innovative methods

It was determined that the best performing spatial model was to consider only k nearest neighbors within a fixed radius while discounting the weights of those neighbors that are farther away. Apart from consistently outperforming other methods, there was some logic to these choices. A problem was observed for neighborhoods in high population density areas. The issue is that success rates are infrequently observed. Only 1 in 1000 ads are clicked on at best,

meaning success rates are low. Clustering causes the spatial interpolation to flatten into less useful results. Limiting spatial points to the k nearest neighbors prevented the clustered areas from out-weighting outlier areas where high success rates are observed. When high levels of clustering were observed, the predicted performances were underestimated due to the high number of poorly performing points. The problem is that the high concentration of poorly performing points overwhelmed the sparse, good performing points in these areas. This method forces spatial elements to only consider congruent and second neighbors. The reason for having a fixed radius is for when a digital zip is in a rural area. By having a limited radial distance, points excessively far away from a digital zip would not be spatially considered even if within the k nearest points. Lastly, weighting was discounted based on inverse distance between points. As the distance increased between two points, the spatial impact decreased. This allowed closer points to make a larger impression on the spatial weighting of a point. This model was used for the remainder of testing. Figure 7 shows a boxplot of the results from 4000 consecutive random backtests. The columns from left to right are a random control, the spatial model, and the theoretical maximum. The y-axis represents whatever success metrics were available for the random campaigns chosen. This included click, view through and pixel rates.



Random	Spatial	Max
.0248	.0259	.0346

Figure 7: Results from backtest

It was observed that compared to randomly choosing neighborhoods, the spatial model had a greater average performance by 4.4%. This meant that the spatial model was able to select new campaign locations that yielded 4.4% higher success rates. Compared to randomly picking locations, taking advantage of the spatial characteristics was able to consistently demonstrate better use of ad space. In 90% of iterations, 1/3 of optimally identified locations were contained in the maximum case. This meant that if the algorithm identified 9 new neighborhoods to advertise in, 3 of them were actually later observed as the highest performing locations in the entire possible location set. The overall low average is due to low starting spatial correlation and clustering of data points. Lastly, it was noted that outliers were never observed below the mean of the spatial model. The cross-marked outliers consistently outperform the rest of the boxplot. This meant that although the spatial model did not always outperform random selection, the outlier points were almost never bad performers, but rather observed exceptionally good results. This meant that the model could consistently identify strong performing outliers, and consistently avoided any incredibly poor performing points.

8 Live Test Results

A similar method to the backtest was used to implement the spatial method into a live campaign. Instead of removing data, the campaign was analyzed to determine which digital zips would perform best that were not already being delivered ads. Only a single iteration was performed in which the spatial regression added three new digital zips to the campaign. The results were initially discouraging as 2/3 of the identified digital zips were low impression areas. This meant that the areas did not necessarily perform poorly, but that people in these neighborhoods did not tend to visit websites that contained places to purchase ad space. This meant that there were minimal opportunities to evaluate the effectiveness of an ad since there were few ads distributed. Although, for campaigns where impression volumes were not an issue, new digital zips tended to outperformed the initial data set when averaging over the same time intervals. Comparing performance over the same days showed that the spatially chosen points on average performed 56.8% better than the original data set. Overall, the average click through rates (CTR) of the campaign was increased by 1.18% over 4 days from the addition of 3 digital zips, meaning the average of the new locations ranked in the 92th percentile for performance

overall. Figure 8 shows the performance of the overall campaign compared to the spatially chosen points.

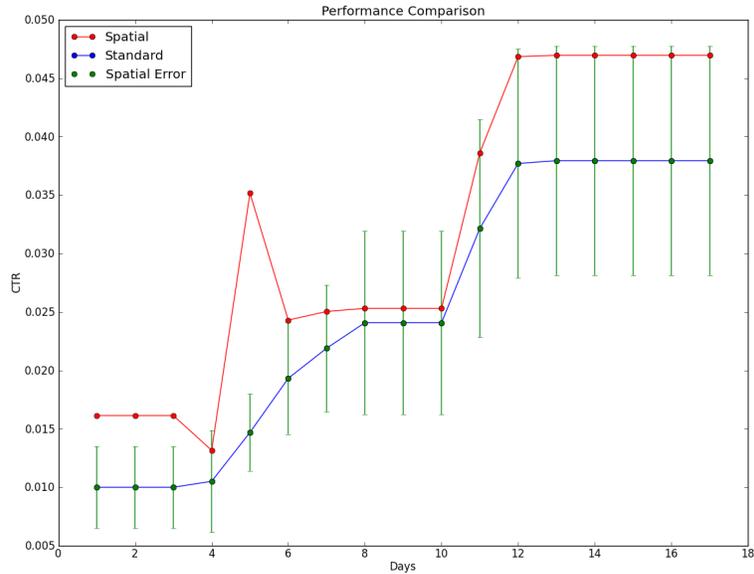


Figure 8: Spatial Points vs. Entire Campaign

Additional tests were performed on two Target and AT&T campaigns using click through rates. This time six new digital zips were added to each campaign. The performance of the spatially added points and the original campaign were monitored for about two weeks. The key observation in each of these tests is that the spatial points are always better than the campaign as a whole. This was a consistent trend for tests performed on other live campaigns. Plots of the performance of these examples can be seen in Figures 9 and 10. In these two examples the original data had 5.61% and 9.8% spatial correlation.

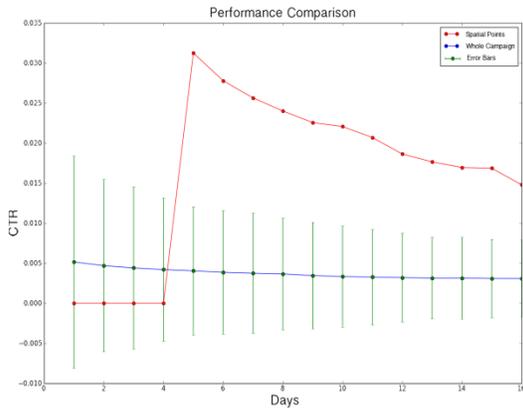


Figure 9: Adding DZs to a Target campaign

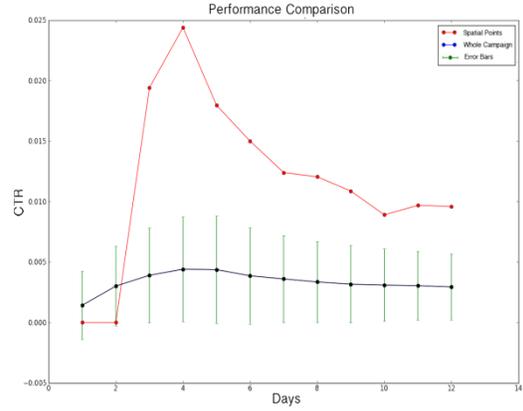


Figure 10: Adding DZs to an AT&T campaign

Just as points can be added to a campaign, digital zips can be identified as poor spatially performing. By removing these points the campaign can be further optimized. Figure 11 shows the results from removing digital zips from a campaign.

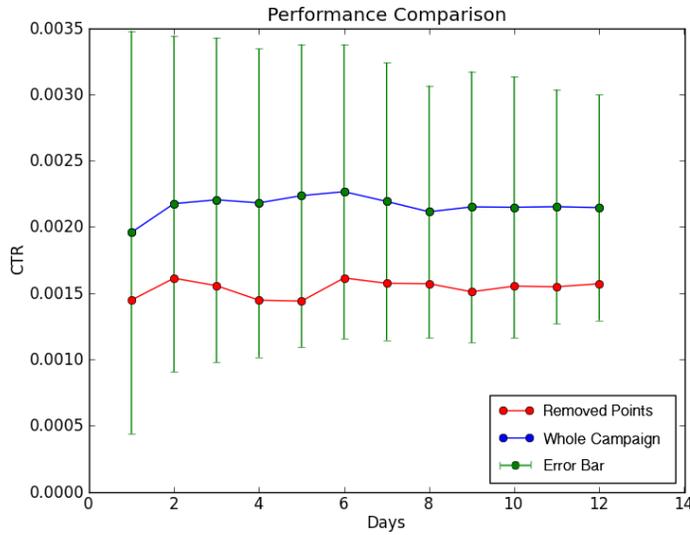


Figure 11: Removing digital zips for Target campaign

The red line is the click through rate for the spatial points selected for removal. The blue line is the overall campaign average with green lines being one standard deviation. The key is that the red line is consistently below the average. By removing these points, the poorly allocated impressions are redistributed to the higher performing areas. Combining the methods of spatially adding and removing points yields the results seen in Figure 12.

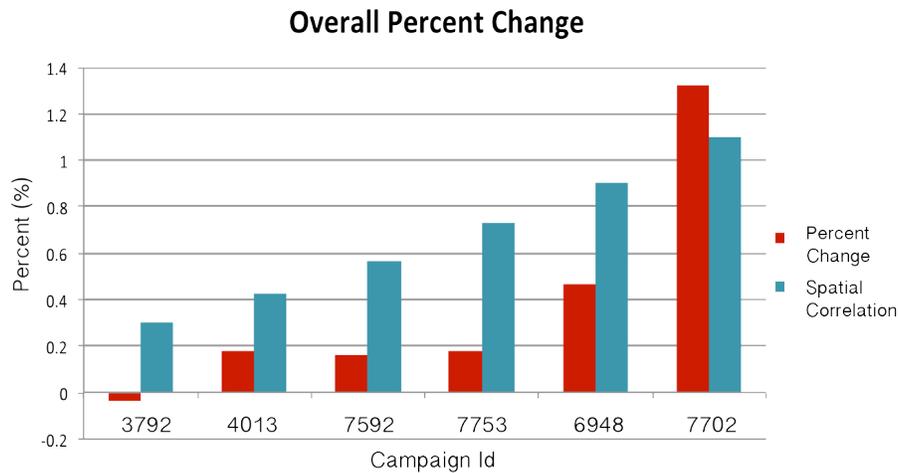


Figure 12: Comparing percent change in performance to spatial correlation

The red bars represent the percent change in overall campaign performance from the changes made by the spatial algorithm. The turquoise bars represent the overall spatial autocorrelation in the campaign. The scale of spatial correlation is 1/10 of the measured value to keep the scale of the superimposed data similar. The campaigns graphed were chosen based on their measured spatial correlation to demonstrate the trend. The success metrics were mixed among campaigns. The specific ads in order were for a cereal bar company, Target, unknown, a local weather station, Target and AT&T. The key is that there is a proportional trend between increasing spatial correlation and percent increase in campaign performance.

9 Known Issues/Comments

The purpose of this section of the paper is to outline the specific issues associated with applying spatial models to the particular situation of using advertising data. These issues may not exist in other data sets, but caused enough trouble in the context of advertising to warrant review.

9.1 Measurement Metric

The main problem with the measurement metric is the scarcity of observed success rates, in particular clicks. A campaign must have a large number of impressions to have a true estimate of the success rate for a given digital zip. It is important to understand that an area with 1 click and 1 impression does not mean that a CTR of 100% is the true value. There must be a threshold number of impressions needed before a success rate should be even considered. In addition to this, performance of the overall algorithm was increased when implementing a decay function that punishes digital zips with lower impressions. For the live and backtested results, a threshold of 500 impressions was used. In the future, this could instead be a dynamic number or function of how long a campaign has been running. Three different mechanisms for discounting weights due to inadequate impression volume were considered. These were:

- Linear decay
- Log-normal decay
- By Michael Els' function:

$$\frac{1+16 \left[\frac{e_i(i_i - e_i)}{i_i^2(i_i + 1)} \right]}{e_i + .01} \text{ where } e_i \text{ and } i_i \text{ are the } i^{\text{th}} \text{ event and impression respectively}$$

The particular decay function chosen was the cumulative distribution (DDF) of the log-normal distribution. The impressions were distributed across a CDF and their respective weights were multiplied by the corresponding areas under the curve. This meant that the weights of areas with relatively low impressions were discounted accordingly. The reason for this is to account for instances where an area is delivered a single ad resulting in one click through. This would indicate a vastly overestimated success rate of 100%, causing the algorithm to further target this

point. The problem is fixed by discounting the success metric seen in areas with low ad volumes. It was observed that without this feature, the same end locations were selected, but that it required many more iterations and observations to hone in on the optimal locations. Without this extra element, it took on average two days for the algorithm to identify when a location did not have enough impression volume to justify a high success rate. When using this additional discounting method, the average time was reduced to an average of 5 hours.

9.2 Low Spatial Correlation

The amount of improvement from using a spatial algorithm is limited by the initial amount of spatial autocorrelation. If the data being spatially analyzed have a low degree of spatial correlation, it is unrealistic to expect vast improvements in the observed success rates for newly added points. This makes some campaigns better candidates than others for spatial analysis. Determining which are best can be done through measuring Moran's I, Geary's C, and the presence of clustering. It was found that high values of Moran's I are a direct indication that a data set has strong potential for increases in performance. The converse of this is not true. Low levels of Moran's I did not mean spatial correlation was not appropriate. When Moran's I was low, Geary's C acted as a good second indicator of spatial correlation. When both of these test yielded poor results, spatial correlation is not likely to have a strong benefit to a campaign.

10 Discussion of Results

Spatial modeling is a unique tool that has a vast range of applications. The general problem with its use is that the results are independent of the path. This paper has demonstrated how such a model can be applied to better predict future outcomes based on past errors. It is not yet clear though as to what this model implies about the system it evaluates. There is great room for interpretation as to what the presence of spatial autocorrelation means for a population in terms of how they respond to advertising.

A large concern of online advertising is that the majority of successful clicks are due to a small subset of the community that does not represent the population as a whole. Success

statistics assume that the population as a single entity is responding to advertising to comprise an average. The problem with this, for instance, is that the rate at which a 35 year old mother clicks an ad is much higher than her son, who has become 'ad blind' from being overexposed to online advertisements. It is clear that not all ads are the same, so the hope is that their responses should not be the same. Although it seems like a basic observation, the same group of individuals does not respond to all advertising similarly. It was observed that where one ad performs well does not necessarily indicate the success of another ad. This is an excellent observation as it indicates that the ads are being responded to by different groups of people rather than the same individuals each time. This means that the specific ads are hitting their intended targeted populations. This leads to the next observation that a click is not critical.

For a campaign such as the Buffalo Local Weather station, clicks have great significance. This ad was distributed in the Buffalo area with the intent of driving additional views to their homepage, which provides local news and weather information. This campaign was very successful, resulting in a high overall click through ratio. The locations that performed the best were the more rural parts of Buffalo City, Wisconsin. This was not an expected result, meaning that the spatial model was very useful in adjusting the performance of this campaign. For other ads, such as for Audi, there was a much lower click through ratio. Although, when tracking delivered ads and Audi home page views, there was a strong correlation between the ad delivery time and the time at which additional views on the home page from that area were made. This means that although the consumer is not clicking on the ad, they are still visiting the webpage from a new tab in the browser. Although this accomplishes the purpose of the ad, it leaves no directly measurable footprint by which the ad performance can be evaluated. Success can only be inferred. Despite strong spatial correlations with Audi ownership, the ads had a mostly uniform response. This meant that it was better to optimize this campaign based on which consumers are most likely to purchase an Audi rather than optimizing on areas just likely to click on an Audi ad. Campaigns for specific consumable goods on the other hand struggled at getting any measurable attention. For instance, a campaign for Kashi Cereal snacks observed almost no clicks. This is likely because it is a familiar consumption item that has minimal online interactions to drive someone to seek more information by clicking on an ad. On top of this, many companies are unwilling to share sales data. This makes it very difficult to identify the effectiveness of advertising for particular products and companies. In essence, this adds an

artistic element to optimally distributing advertising based on optimizing to achieve the desired result. Apart from these observations, the final algorithm can be used to make further observations based on ad performance.

By virtue of observing successful results, it is clear that individuals are arranged across the country in a fashion that maintains spatial correlation to their neighbors. Although, further insight into this arrangement can be made based on what parameters optimized the final solution. Apart from generating an actual spatial algorithm to take advantage of inherent spatial autocorrelation, this solution can be back solved to draw broader conclusions on these evaluated areas. A more obvious, but relevant example is the restriction of nearest neighbors to staying within a fixed distance radius. If there were no limit, it was found that bigger cities such as Atlanta and New York were strongly correlated in their response to similar ads. The problem here is that this is not a spatial relationship, but rather a big city similarity. The responses have nothing to do with where those cities are located but just by essence that they are both large cities. Conversely, if only first neighbors are considered, there is not enough quality quantitative data within such a small radius to fairly evaluate the spatial correlation between small neighborhoods. Through many iterations it was determined that a 30 mile radius was best. This is large enough to consider entire neighborhoods and cities without having the problem of considering effects larger than spatial autocorrelation. On this same note, 30 miles is quite a large radius for many areas, where multiple completely different neighborhoods get incorrectly grouped. This is why within this radius it was a good idea to use inverse proportional weighting against distance.

Another interesting topic is clustering. In regions of high density populations, there are many digital zips clustered very closely together. These zips are constructed based on the distributions of people that live there. In a place like New York City, these areas are almost on top of each other and respond entirely differently to various ads. An ad for a Target clothing line saw great success rates where there were virtually no responses an eighth-mile away at the next digital zip. This is an indication that there may be strong socioeconomic divides that dictate consumption habits. Although, ads for free services, such as an online weather website, exhibited a much more uniform response to ad exposure. The notion of 'free' is a common trend observed in almost all areas of economics. This driving force tends to cause individuals to change

consumption habits more drastically than with just price reductions. There is also likely some element of attachment to a local product over one that is catered by a chain company such as Target. This is especially true of a city such as Durham, North Carolina. Companies like Local Yogurt and Bull City Burger perform exceptionally well due to this local attachment.

Overall, the original, non-spatial prediction adequately evaluated which locations would have the best performing ads. The introduction of the spatial element was most useful for correcting the model where large amounts of error were observed. This was especially useful for adjusting campaigns that misevaluated the intended consumers. In these cases, the initial observations of ad performances leave a lot of room for spatial improvement through many iterations of the algorithm. After this analysis it is evident that there are multiple instances where predicting the outcome of advertising is akin to shooting clay targets in the dark. Unless time and location stamped sales data is provided, the absolute optimal solution is nearly impossible to attain.

11 Conclusion

This document reviewed spatial modeling and possible methods for improving neighborhood selection based on observed performance information. It was demonstrated how spatial modeling can be used to optimally adjust the locations in which an advertiser should allocate resources. The ultimate purpose is to better identify and target areas of high success rates in order to improve campaign performance. The only problem with these techniques is that although the model is improved, no additional knowledge is gained. All that is gleaned is that there are underlying, unexplained spatial relations in the data. It is not known why this is observed. Basically this means that the process of spatial regressions is path independent. In context of this problem, there are two evaluations of this concept. This first idea is that as the original model becomes more sophisticated, the need for spatial consideration should diminish. This means that every time an additional parameter is added to the initial, non-spatial regression, the strength can only increase. As the number of parameters evaluated approaches infinity, the strength should theoretically reach infinity and optimally predict the outcome. This would be very difficult and impractical. Likewise, there will forever be noise in the form of unpredictable

error due to completely random events. No matter how comprehensive of a model, it can never be perfect. The theory breaks down when trying to evaluate non-quantitative parameters. This leads way to the second concept of path independence of the spatial model. Spatial considerations can still remain significant when hard to measure elements, such as psyche or simply preferences, are confounded as noise in the data. Further research for intellectual purpose would be to consider spatially regressing the explanatory variables in the original model rather than the observed outcomes alone. Although this has no direct benefit to an advertising agency, this would provide insight as to how some variables spatially correlate to a success rate. Although less likely to occur, it would also be valuable to convince retail points to release sales data to definitively determine the success of a campaign.

12 References

Anselin, L. (2005). *Spatial Regression Analysis in R*. Urbana, IL: Center for Spatially Integrated Social Science.

Bivand, R., Pebesma, E., Gomez-Rubio, V. (2007). *Analyzing Spatial Data in R: Worked example: spatial autocorrelation*. Norway: Springer Science and Business Media, LLC.

Brusilovskiy, E. (2009). *A Brief Introduction to Spatial Regression*. Philadelphia, PA: Business Intelligence Solutions.

Diggle, P., Fuentes M., Gelfand, A., Guttorp, P. (2010). *Handbook of Spatial Statistics*. Miami, FL: CRC Press, Taylor & Francis Group.

Gaetan, C., Guyon, X. (2010). *Spatial Statistics and Modeling*. New York, NY: Springer Science and Business Media, LLC.

Lembo, A. Jr. (2008). *Spatial Autocorrelation: Moran's I and Geary's C*. Salisbury University.

LeSage, J., Pace, K. (2009). *Introduction to Spatial Econometrics*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Wash, P. (2010). *InvSWM_F and nnSWM*. Matlab File.