A GEOSTATISTICAL ANALYSIS OF HOUSING PRICES IN FARGO

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Tika Ram Lamitare

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

April 2017

Fargo, North Dakota

North Dakota State University
Graduate School

Title

A GEOSTATISTICAL ANALYSIS OF HOUSING PRICES IN FARGO

By

Tika Ram Lamitare

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Gang Shen
Chair

Dr. Seung Won Hyun

Dr. Lei Zhang

Approved:

May 2, 2017
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

This study investigated housing prices by applying geostatistical regression techniques to identify the significant factors affecting residential housing sales prices in Fargo, North Dakota**.** The study used a subset of residential housing price sales data for the year 2015. The study found moderate spatial dependency among properties. Some of the statistically significant variables were found to be age, total area of property, number of parking spots, air conditioner and the status of basement finish. Finally, predictions on new locations were made based ordinary linear regression model and regression kriging technique used for the geostatistical models.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Gang Shen, for his advice and guidance. I want to especially thank him for clearing my doubts related to geostatistical modeling, which in turn allowed me to explore the beauty of spatial statistics. Also, I would to express sincere thanks to my committee members, Dr. Seung Won Hyun and Dr. Lei Zhang, for providing me valuable feedback. Finally, I am very grateful to Ben Hushka and James Haley for providing me the data and the information about the variables used in the data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

With a strong economy and labor force, Fargo has experienced economic growth in the last few years. This economic growth has led to increases in real estate transactional activities. An example of increase in transactional activities was reported by Bishop (2016) on a South Fargo property whose value increased by $63000 in one year. Housing prices are heavily reliant on the core structural characteristics and neighborhood characteristics of properties. Some structural characteristics would include age of a property and its area, total number of bedrooms as well as bathrooms, type of air conditioner, and other similar characteristics. Neighborhood characteristics would include proximity to a shopping mall and highway, closest school district, access to nearest recreational facilities, crime rates, and other similar characteristics. A combination of the above factors and potential external factors ultimately determine property values.

All the housing sales transactions in Fargo are reported to the City of Fargo. The City of Fargo also has its own methods to assess property values. The City (n.d.) defines its appraisal technique "as …the systematic appraisal of groups of properties as of a given date using standardized procedures and statistical testing." One common feature of mass appraisal technique involves the comparison of properties with similar characteristics in assessing the value of a property that shares similar characteristics with those properties that were compared. Besides sharing similarity in their structures, houses have their own geographical locations, which make spatial dependency an issue in determining property values.

The question arises: does location matter? Ordinary linear regression modeling has traditionally disregarded the need to incorporate the location structure, but as Tobler's Law states, "Everything is related to everything else, but near things are more related than distant things." The issue of spatial proximity lends itself towards geostatistical modeling. While the role of external

covariates in determining the prices of properties is apparent, it is also equally imperative to incorporate the role of locations by using them as a function of distance. Doing so takes into account the existence of spatial autocorrelation structure, thus making the task of regression modeling applicable in modeling housing prices as well as more effective.

So the main objective of this thesis is to create a regression model that quantifies the impact of spatial proximity by detecting the existence of spatial autocorrelation structure. While accomplishing that, this thesis starts with the classical method of variogram modeling and furthers into the parametric method. The next aim of this thesis is to identify the significant factors that impact the residential housing prices in Fargo. Finally, this thesis uses regression kriging technique to predict values for new houses at new locations. The application of geostatistical techniques in modeling housing prices, as demonstrated in this thesis, should help government officials, market research organizations and realtor groups in their approach to property valuations.

2. DATA DESCRIPTION

The City of Fargo's Accessor Office annually collects data on the houses that are sold in a given year. This thesis uses a subset of 2015 property sales data in Fargo. At first, geostatistical models with exponential and spherical covariance functions were trained on a dataset with 1352 observations. Finally, predictions based on regression kriging technique were performed on a dataset with 341 observations. The split ratio has been approximately 80-20, with 80% of the data on the training data set and the remaining on the testing data sets. The dependent variable, price, was transformed to log scale. The property type variables had four levels: single family, duplex, three plex and twin houses. Duplex, three plex and twin houses were jointly considered as non-single family dwellings, thereby creating only two groups, single and non-single, for property type. Style of the house, as denoted by story height on the data, had many different levels. Some story height levels were combined based on their similar attributes. For instance, story heights with levels of one story and one and half story were considered jointly as one story. Similarly, bi-level story height and bi-level with additional were jointly considered as a bi-level story height. The categories of story height included bi level, one story, split level and two story. Air conditioner variable had three groups: central, wall and none. The 'none' category in a given property implies that the property does not have built in air conditioner.

The data also had information on basement finish, which the city classifies by none, 25 %, 50%, 75%, and full finished. These different levels were transformed so that none and 25% were considered as less or a quarter finished, 50% and 75% were recoded as half or more than half finished and full finished was left as it originally was. The types of garages were also included in the analysis. There were five types of garages, namely attached, built-in, combined, detached and none. Built-in, combined, detached and none garage types were jointly considered as non-attached

garages, due to which the garage type variable had two levels: attached and non-attached. Information on flood history, denoted by X_100YrFlood variable had two categories: yes and no. Properties that were deemed to experience flooding in a given year by the city had 'yes' entries, while the properties that were not deemed to experience flooding had 'no' entries.

The city also recoded numerical entries in a categorical way, such as number of bathrooms was coded using "1" for 1 bathroom, "2" for 1 and half bathrooms. Instead of using bathrooms through dummy coding, houses with one and one and one half bathrooms were considered as one, houses with 1.75 ,2, 2.5 bathrooms were considered as two, three and three halves were considered as three, four and four halves were considered as four. Additionally, by calculating the difference in longitude of the intersection at Main Ave and 25th St S to the longitude of houses that were used for modeling and prediction, a new variable, denoted by difference, was created. The variable, garages, refers to the number of parking spots inside a garage. All of the numerical predictor variables were standardized. Furthermore, geographical coordinates latitude and longitude were transformed into UTM Northing and Easting. It has to be noted that the geographical coordinates were not used as predictor variables.

A list of the variables used is given in table 1 and a sample of five raw data is shown in table 2. The map on figure 1 shows the concentration of properties that have been used for training the model. The map, as well as other maps used for analysis, was generated using *ggmap* function in R (Kahle and Wickham). Based on the figure, it seems that residential properties in Fargo are heavily concentrated in three major regions. The most heavily concentrated region lies around the area bounded by South University Drive, 32nd Avenue South and 45th Street South. Also the other heavy concentration of residential properties in Fargo is around the boundary of 13th Avenue S,

25th Street S, Main Avenue S. Finally, the other heavily concentrated area is the North Fargo region, including main Avenue N, N University Drive and towards 19th Avenue North.

Table 1: Variable Description

Variables	**Role**
Price (log transformed)	Dependent variable
Property Type	Categorical
Story Height	Categorical
Segment Square Feet	Numerical
Building Segment Feet	Numerical
Air Condition	Categorical
Basement Finish	Categorical
Number of Bathrooms	Numerical
Garage Type	Categorical
Number of Garages	Numerical
Flood region	Categorical
Age	Numerical
Number of Bedrooms	Numerical
Difference	Numerical

Table 2: Sample Data

Lon	Lat	lnprice	age	SegSqFt	BldgTotSF	bed	baths	garages	proptype	storyheight	Aircond	basementfinish	garagetype	Flood
665482	5187860	12.4288	20	10471	1126	4	2	3	SF	BL	Central	Fullfinished	Attached	Yes
665667	5191529	12.3779	40	9600	1436	4	2	2	SF	BL	Central	Fullfinished	Attached	No
666747	5192497	11.8776	61	6600	864	3	1	2	SF	OS	Central	Quarterorless	Notattached	No
666655	5192285	11.9505	60	5964	1098	3	2	2	SF	OS	None	halformore	Attached	No
666992	5190876	12.0347	56	9375	1056	3	2	1	SF	OS	Central	halformore	Attached	No

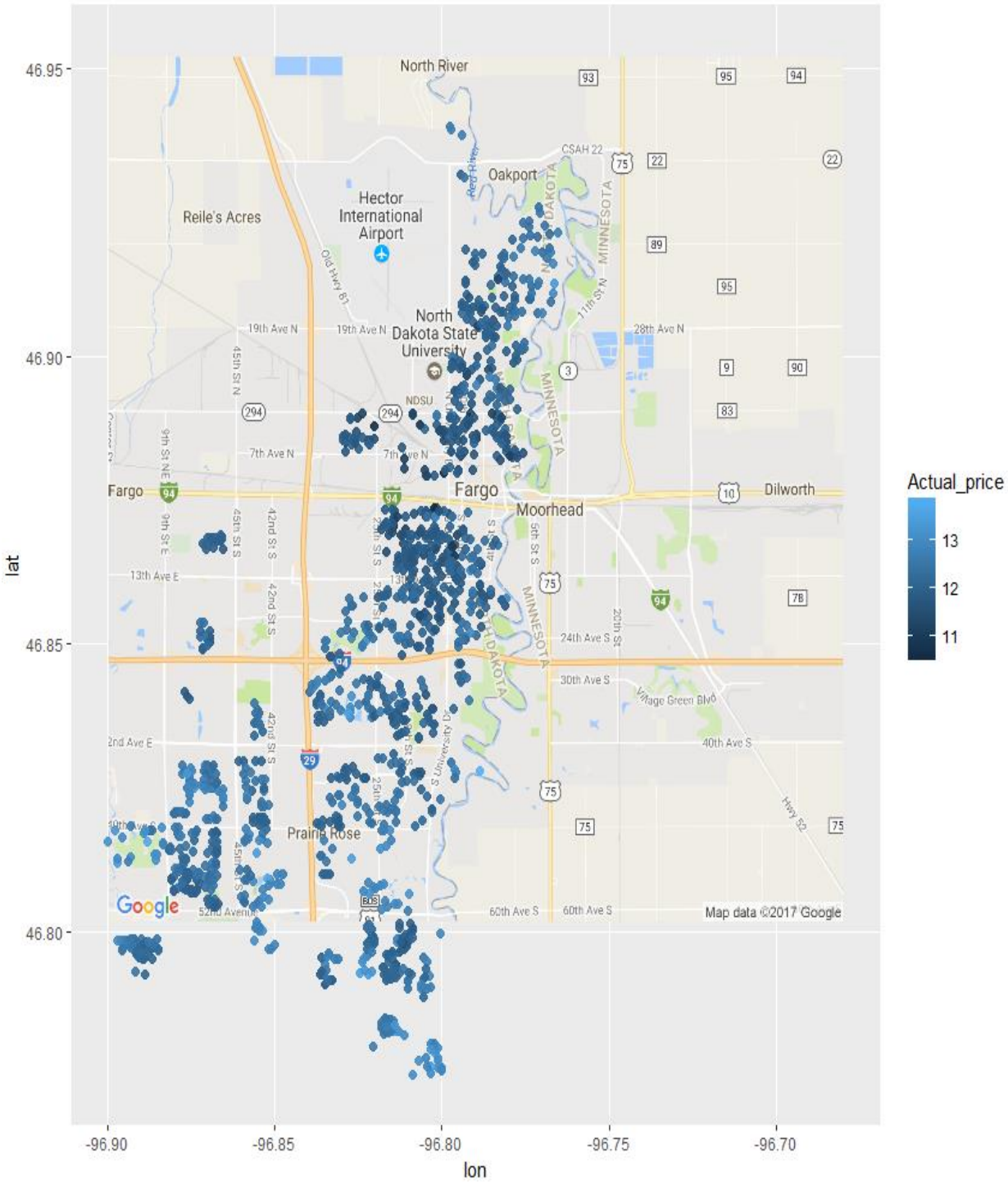

Figure 1: Plot of Houses with their Values

3. LITERATURE REVIEW

Chica-Olmo (2007) cited earlier research while categorizing housing price regression modeling into two categories: hedonic regression modeling and spatial regression modeling. Econometricians have usually referred to housing price regression models as 'hedonic regression models.' Hedonic regression models emphasize the importance of structural characteristics of a property, as well as neighborhood characterstics of the property in determining property values... Koramaz and Dokmeci (2012, p.1235) found size of a property, "centrality, accessibility and distance to the coast are spatial determinants found to be statistically significant" in regards to housing price models in Istanbul, Turkey. He et al. (2010, p.923) also concluded that the "distance between the housing and the downtown area, floor area ratio and land transaction price" to be statististically significant and concluded them to be important factors for housing prices in Beijing. While the variables considered in both of these models were not all the same, a heavy emphasis was placed on neighborhood characteristics of a property.

Like the hedonic regression models, spatial regression models or geostatistical models place an equal importance on the core characteristics of a property, but unlike the former, the latter considers that the location of a house can be used to model the spatial dependence between houses. Laying out the need for a spatial regression model, Chica-Olmo (2007, p.91) referencing Dubin's study, stated, "Usually housing sale price will be directly related to the sale price of other neighboring houses. Location is probably the most important variable used to explain house price. Spatial autocorrelation is present when location is very important to housing prices." Bourassa's et al. (2005) research on residential housing values in Auckland, New Zealand drew two significant conclusions regarding the impact of spatial autocorrelation. In terms of overall model predictor, geostatistical model performed better than the OLR model but when neighborhood dummy

characteristics were added to the OLR model, the OLR model performed better. Therefore, it cannot be said that geostatistical models always perform better than the OLR models. The choice of using one model over another may strictly depend on the availability of auxiliary variables and the degree of spatial dependency. But this thesis primarily aligns with Chica-Olmo's assertion regarding the particular location of house and its usefulness in determining the underlying correlation structure among properties.

While the hedonic and geostatistical modeling techniques have primarily relied on the core structural and neighborhood characteristics, Dubin's (1998) work on real estate prices in Baltimore provides another interesting perspective. Dubin's research had three different modeling components: an OLR model, an OLR model with the inclusion of location coordinates and a geostatistical model with the inclusion of location coordinates. Dubin found the geostatistical model performing better than the OLR models with and without location coordinates. Dubin also concluded size of the house, number of bathrooms, number of bedrooms to contribute positively towards increasing the property prices, while increase in the age of a house was found to impact negatively on the residential housing properties.

Gelfand et al. (2004) studied on the two regions, Highland and Sherwood, in the City of Baton Rouge, Louisiana. Their study found age, living area and area encompassing patios, garages and carports to be statistically significant factors for determining prices in both regions. The number of bathrooms was found to be statistically significant only in the region of Sherwood. Their study relied on geostatistical technique. In terms of spatial autocorrelation, the value of range was found to be around 2 and 1 km for the respective regions, probably a clear indication of weak autocorrelation.

4. METHODOLOGY

4.1. OLR Model

Although geostatistical methodology is the key to regression model building in this thesis, at first, a multiple regression model was created with the following form:

$$\log(y_S) = \beta_0 + \sum_{j=1}^{17} \beta_j x_{sj} + \varepsilon_s, \text{ s=1 to 1352} \quad (4.1)$$

where $\log(y_i)$ is the log-transformed dependent variable price, $x's$ are the independent predictor variables, and the regression coefficients were estimated in the following way:

$$\beta=(X^TX)^{-1}X^TY. \quad (4.2)$$

The model above can be classified into the deterministic trend and the random component. The random error component, ε_s is assumed to be normally distributed with the following two properties:

- $E(\varepsilon)=0$
- $Var(\varepsilon)= \sigma^2 I$

, where I as an identity matrix. When similarities emerge in terms of spatial location, Abraham and Ledolter (2006, p.127) commented that "errors for measurements taken in close spatial proximity are correlated," thereby indicating a need of a better modeling technique in place of OLR model. The discussion of the error component is important here due to some geostatistical assumptions that will be discussed in the next section. In the geostatistical process of detrending, residuals were first extracted from the OLR model discussed above.

The interest in OLR model became apparent as Diggle and Ribeiro (2007, p.100) mentioned using "the residuals to identify a suitable parametric model for the covariance structure

and to obtain initial estimates of covariance parameters." The justification for getting residuals and its use in the geostatistical model building has two main reasons:

- The location of a house is particularly important in model building process, but equally important is the fact that external covariates can influence the price of a house. The process of residual extraction is the process of detrending nonstationary components.
- Residuals represent the stationary component for the spatial process under study.

4.2. Geostatistical Model

Let us assume that s represents spatial coordinates x and y, where x represents the UTM longitude and y represents the UTM latitude. Then the geostatistical model becomes

$$\log(y_s) = \beta_0 + \sum_{j=1}^{17} \beta_j x_{sj} + v_s + \varepsilon_s, \quad s = (\text{UTM Northing, UTM Easting}) \qquad (4.3)$$

, where the deterministic part is came as outlined in equation 4.1. The second term, v_s, relates to the spatial error and is assumed to be stationary and isotopic with the following conditions:

- $E(v_s)=0$
- $C(v_s, v_{s+h}) = \tau^2 \rho(||h||; \theta)$
- $C_v(h) = \tau^2 \rho(||h||; \theta)$

, where h is a separation lag vectors between two phenomena under study, τ^2 is the variance of the spatial process or partial sill, $\rho(||h||; \theta)$ is the correlation function chosen from a family of isotropic covariance functions, $||h||$ representing the Euclidean distance between two pairs and θ is the range parameter. The covariance between two pairs have been defined as a function of distance between them and not the location, thus covariance stationary condition holds. The final term in equation (4.3), ε_s is related to the non-spatial error and that can be attributed to a number of factors such as

measurement errors or differences in properties that are located in closer proximity. The non-spatial error term, ε_s, is normally distributed with a mean of zero and variance of nugget, i.e. $\varepsilon_s \sim N(0, \sigma^2)$. This thesis has used τ^2 in place of σ^2 to relate it to the OLR model outlined in equation 4.1. Therefore, τ^2 relates to the spatial error part and σ^2 relates to the non-spatial error. So τ^2 is defined as partial sill variance, whereas σ^2 as nugget variance. While geostatistical modeling techniques incorporate different covariance parameters based on the chosen covariance functions, the covariance functions used in this thesis, exponential and spherical, relies only on σ^2, τ^2 and θ. The last parameter, θ , is the range parameter and its use in the geostatistical modeling will be clear in the later sections.

4.3. Variogram

Variogram plots are needed to quantify the underlying correlation structure. Experimental variograms are plotted from the raw data, while theoretical variogram provides the mathematical basis on covariance parameters that can be extracted from a plotted experimental variogram. Theoretical variogram can be defined as:

$$\gamma(h) = E[(Y_s - Y_{s+h})]^2 / 2. \quad (4.4)$$

Solving equation (4.4) allows to make the following conclusions:

$$\gamma(h) = C(0) - C(h) \quad (4.5)$$

, and as h tends to infinity, then

$$\gamma(h) = C(0) = \sigma^2 + \tau^2. \quad (4.6)$$

Based on the discussion above, it can be concluded that as distance tend to increase the variogram reaches the sill value, which is also the variance. Furthermore, at a smaller separation distance, the theoretical expectation is that the variogram value should be 0. But due to some measurement

errors or dissimilar feature among properties that are in closer proximity, the variogram value is greater than 0. When the variogram value is greater than 0, that value is defined as nugget variance.

Above all, computation and visual exploration of the experimental variogram is the key to building a parametric model. The experimental variogram computation was based on the OLR model residuals and can be shown as:

$$\gamma_{(h)} = N_h \sum_{i=1}^{N} (\hat{e}_{s+h} - \hat{e}_s)^2/2, \quad (4.7)$$

where h represents the Euclidean distance between two pairs of observations. For n number of observations, then the total number of distance pairs is $\frac{n\cdot(n-1)}{2}$. Furthermore, steps in the computation of variogram can be explained in the following ways:

- find Euclidean distance between each distance pairs
- find the squared difference between $\hat{e}_s$ and $\hat{e}_{s+h}$ for each s and s+h
- calculate semi variance using the equation (4.7), where N_h represents number of pairs in each squared difference calculation
- Based on the choice of lag, also find the average of semi variance values based on the distance pairs that are in between the lag distance
- Then an empirical variogram can be fitted with average distance on the x-axis and semi variance values on the y-axis.

Based on the completion of these five steps, an empirical variogram can be fitted. Thus, empirical variogram guides towards parametric modeling since it allows to explore possible covariance functions for the covariance parameters, namely σ^2, θ and τ^2. While σ^2 and τ^2 have been already defined above, θ is the range parameter and it represents the value on the x-axis when the variogram nearly reaches the sill on the y-axis. The interest in geostatistical modeling is within the

origin to the value of θ. Beyond θ, autocorrelation seems to get weaker as separation distance increases.

It should be mentioned that the computation of experimental variogram and its plot have been used to identify the initial estimates for the covariance parameters. The experimental variogram is heavily subjected to binning procedures, therefore the application of parametric method has been used in this thesis. The discussion of covariance parameters based on the experimental variogram as well as parametric method is in section 5.1.

4.4. Covariance Functions

Once the experimental variogram was created, the next step was to use covariance functions to model the covariance parameters. There are many types of covariance parameters that could be chosen, but this thesis used exponential and spherical covariance functions. They are the most commonly used covariance parameters. While some have selected covariance functions based on the shape of the experimental variogram, this thesis relies on parametric modeling of the chosen covariance functions.

4.4.1. Exponential Covariance Function (ECF)

One of the most widely used covariance function, its covariance can be listed as follows:

$$C(h) = \begin{cases} \tau^2 e^{\frac{-h}{\theta}}, & \text{for } h>0 \\ \sigma^2 + \tau^2, & \text{for } h=0 \end{cases}$$

and its corresponding variogram can be defined as:

$\gamma(h) = C(0) - C(h)$.

While developing the model in section 4.2, the covariance was defined in terms of its relationship with the correlation function. Thus, ρ (||h ||; θ) = $e^{\frac{-h}{\theta}}$. In the case of exponential variogram, the variogram reaches the sill asymptotically. This indicates that the exponential function does not have a true range beyond which the autocorrelation of the observed phenomenon decays to zero. But its effective range has been defined as θ_1=3θ, so beyond θ_1 covariance starts to decay towards zero.

4.4.2. Spherical Covariance Function (SCF)

The spherical covariance function is given below:

$$C(h)= \begin{cases} \tau^2 \left(1-\frac{3}{2}\cdot\frac{h}{\theta}+\frac{1}{2}\cdot\left(\frac{h^3}{\theta^3}\right)\right), & \text{when } 0<h\leq \theta \\ 0, & \text{, when } h> \theta \\ \sigma^2+ \tau^2, & \text{, when } h=0 \end{cases}$$

and it's ρ (||h ||; θ) = $1-\frac{3}{2}\cdot\frac{h}{\theta}+\frac{1}{2}\cdot\left(\frac{h^3}{\theta^3}\right)$.

Unlike the case of exponential, covariance between two pairs goes to 0 when the distance between them is greater than the range value. Thus, spherical covariance function has a true range, beyond which covariance between two properties goes to 0.

4.5. Parameter Estimation

Now that the variogram has been defined and its relation to the covariance function has been established, the other important task was to create a geostatistical regression model. Different methods have been proposed to fit spatial correlation in a regression model, most importantly weighted least squares method, maximum likelihood estimation and restricted maximum likelihood estimation. Nevertheless, estimation of a spatial regression model in itself is a daunting

task. This is, because, correlation estimates based on the experimental variogram are needed to estimate a spatial regression model. However, the correlation estimates based on an experimental variogram highly depend on the way binning process is done. Therefore, the objective is to use parametric method to estimate the regression model and covariance parameters.

The issue of geostatistical modeling is a statistical problem as well as a computational problem. Different authors have described different geostatistical parametric modeling techniques, but emphasis should be given to the fact that the computational complexity of spatial modeling techniques require an equal understanding of the statistical methodology as well as algorithmic design. Therefore, this thesis primarily considers the work of Diggle and Ribeiro (2010), as it relates to both statistical modeling and algorithmic design. Diggle and Ribeiro are authors of **geoR** package in R, which was used to generate the model. Therefore, their technique of parametric modeling will be discussed below.

Let us first consider the following multivariate normal distribution:

$Y \sim \text{MVN}(X\beta, \tau^2 \rho_\theta + \sigma^2 I)$, $\quad \rho_\theta = \{\rho(||s_i - s_j||; \theta)\}_{i,j=1}$

Where X relates to the covariates matrix with all ones in its first column, β is the regression coefficients that need to be estimated, ρ_θ is an isotropic correlation function and its relation of the covariance function can be described as, $C(h) = \tau^2 \rho_\theta$. Given the distribution above, the log-likelihood is:

$$L(y; \beta, \tau^2, \sigma^2, \theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\tau^2 \rho_\theta + \sigma^2 I|) - \frac{1}{2}(y - X\beta)^T(\tau^2 \rho_\theta + \sigma^2 I)^{-1}(y - X\beta) \quad (4.8)$$

, where || is the determinant. After this, Diggle and Ribeiro (2010) discussed the computational side of maximizing the equation on (4.8). Per their parametrization technique, $W = \rho_\theta + v^2 I$, where

$v^2 = \frac{\sigma^2}{\tau^2}$ and *I* is the identity matrix. Then they found the maximization procedure yields the MLE estimates of the regression coefficients to be the generalized least squares (GLS) estimator. This can be written as:

$$\beta_{GLS} = (X^T W^{-1} X)^{-1} X^T W^{-1} Y \tag{4.9}$$

where W is a n*n symmetric matrix and $W = \tau^2 \rho_\theta + \sigma^2 I$, where I is the identity matrix. Furthermore, to obtain the new covariance and regression parameters that maximizes the log-likelihood, they obtained a concentrated log-likelihood and it can be shown as:

$$L_0 (v^2, \theta) = -\frac{1}{2}\{n \log(2\pi) + n \log \hat{\sigma}^2_{GLS}(W) + \log(W) + n\} \tag{4.10}$$

, where $\hat{\sigma}^2_{GLS} = \frac{1}{n}\{y - X\beta_{GLS}\}^T W^{-1}\{y - X\beta_{GLS}\}$.

Once the regression and covariance parameters are determined by the model in equation (4.9) and (4.10), new estimates need to be given to the *likfit* function until the estimates given by the function and the estimates given to the function stayed the same. The *likfit* function uses numerical optimization algorithm based on minimizing negative log likelihood function to estimate the final parameters.

4.6. Regression Kriging

The steps mentioned above were used to create geostatistical regression models. Now the final goal in any geostatistical analysis is to perform interpolation. Once the covariance and/or variogram parameters have been established and a geostatistical regression model has been created, kriging procedures can be applied to predict data values at new location. Some of the most popular kriging techniques are simple kriging, ordinary kriging, universal kriging and kriging with external drift. While these techniques have their own merits, the application of RK in this thesis is

primarily due to the very reason that it allows deterministic model and residual kriging process to be modeled separately.

According to Hengl et al. (2007), the RK prediction model is done in the following way:

$$\hat{y}(s_0) = \beta_0 + \sum_{i=1}^{p=13} \beta_i x_i(s_0) + \sum_{i=1}^{341} (\lambda_i)^T e(s_i), \quad (4.11)$$

and the prediction variance is:

$$\sigma^2(\hat{y}(s_0)) = (x_0 - x^T W^{-1} g_0)^T (x^T W^{-1} x)^{-1} (x_0 - x^T W^{-1} g_0)^+ \sigma^2 + \tau^2 - g_0^T W^{-1} g_0 \quad (4.12)$$

where $\hat{y}(s_0)$ is the value to be predicted at s_o location. The beta coefficients are derived based on the discussion on section 4.5, $x_i(s_o)$ is the value of predictor variables at a new location. This constitutes the deterministic part of RK. In terms of residual kriging, λ_i refers to the kriging weights that are estimated based on the covariance functions defined in section 4.4 and $e(s_i)$ are residuals from the fitted GLS model. In terms of the prediction variance, x_0 represents the vector of predictor variables at new locations, x is the design matrix of the original data locations, W is already defined in the previous section, g_0 is the covariance vector of residuals at new data locations, and $\sigma^2 + \tau^2$ is the sill or variance.

Kitadnis (2003, p.154) defined the method of SK by stating, "If z(x) is stationary, the mean is a known constant and the covariance is a function of the distance," then the SK estimator becomes:

$$\hat{y}(s_0) = m + \sum_{s=1}^{n} (\lambda_i)\, e_s \quad (4.13)$$

where e_s is a vector of residuals and m is a known mean. SK procedure is also referred as Best Linear Unbiased Estimator as the derivation of its weights based on minimizing $E[(y(s_0)-\hat{y}(s_0)]^2$ creates the following kriging weights,

$$\lambda = W^{-1} g, \tag{4.14}$$

where W is as defined above. For a new location to be predicted, g_0, is developed as a covariance vector of residuals:

$$g_0 = \{C(s_0, s_1), C(s_0, s_2), \ldots\ldots\ldots\ldots, C(s_0, s_{1352})\}$$

, where s_0 represents the new location and the covariance as a function of distance is as defined in section 4.4. Finally, the krig residuals are added back to the deterministic trend of the model as given in equation (4.11).

5. ANALYSIS

5.1. Experimental Variogram

Experimental variogram based on equation (4.7) was first plotted against the separation distance. In terms of formal statistical inference, experimental variogram has been used in the aspect of exploratory analysis, not in terms of parametric model fitting. Diggle (2007, p.104) mentioned, "Our view is that the sample variogram should be regarded primarily as a helpful initial display to identify broad features of the underlying covariance structure of the data, but not as a formal method of parametric inference." Diggle's assertions are applicable in this thesis because the binning process is very subjective and the covariance parameters estimated based on the experimental variogram may change if the binning process is altered.

R package **geoR** was used to perform all of the analysis, including experimental variogram and model fitting. Before plotting an experimental variogram, outliers were detected and removed, new maximum distance was defined as the half of the maximum distance and the plotted variogram at least had 30 pairs in each bin. While using half of the maximum distance and having 30 number of pairs in each bin seem to be the rule of thumb in geostatistical literature, consideration was also given to the outliers. Residual outliers can affect the variogram plot and removal of them is strongly suggested if there is enough reason to do so. Kim's (2015) study found that the model after deleting outliers performed much better than the model that included the outliers. Outliers were checked on a case by case basis and 10 of them were removed.

After removing the outliers and setting up the new maximum distance, the next step was to determine the number of classes of distance that would be used for plotting the experimental variogram. The variogram was plotted using different numbers of pairs. When the classes for distance used was 50, 100,500 and 1000, then each class had more than 30 pairs in each bin.

However, when the classes for distance was increased to 2000, 2500 and 5000, some of the classes had less than 30 pairs in each bin. It is still imperative to remember that the process of variogram plotting was just to obtain some initial estimates for the parametric modeling through maximum likelihood estimation.

The output in table 3 shows the value of initial estimates that was 'fitted by eye'. Based on the table below, the value of range parameter is same in different number of classes that were used. The table below shows the changes in values of sill and nugget, though small, as classes increased. This clearly aligns with Diggle's assertions that experimental variogram should be used for obtaining some initial estimates that can be used for parametric modeling. In terms of parametric modeling that will be discussed in the next section, initial values of covariance parameters for 100 classes were used.

Table 3: Comparison of Covariance Parameters for Different Classes

Parameter	Classes	Value
Sill	50 CLASSES	0.036
Range	50 CLASSES	3000
Nugget	50 CLASSES	0.005
Sill	**100 CLASSES**	**0.037**
Range	**100 CLASSES**	**3000**
Nugget	**100 CLASSES**	**0.005**
Sill	500 CLASSES	0.040
Range	500 CLASSES	3000
Nugget	500 CLASSES	0.005
Sill	1000 CLASSES	0.044
Range	1000 CLASSES	3000
Nugget	1000 CLASSES	0.0055

5.2. Fitted Covariance Parameters

In section 4.5, the method of parametric modeling for geostatistical models were discussed. Using the method of MLE, the generalized least squares estimator of the beta coefficients were found to be the GLS estimators.

But since the elements of W were not known ahead and were estimated using the initial estimates discussed in section 5.1, the generalized least squares estimators of the regression coefficients simply become estimated generalized least squares estimator:

$$\beta_{EGLS}=(X^TW^{-1}X)^{-1}X^TW^{-1}Y \tag{4.15}$$

, where the final elements of W were estimated using initial values and reassigning new values until the assigned values and the new values returned by the *likfit* function stayed the same.

Once the covariance parameters of W matrix were estimated and deemed to be final, the values in table 4 below were used to populate the W matrix using covariance functions discussed in section 4.4.1 and 4.4.2 and with partial sill=.014, range=481.8 meters, effective range=1445.4 meters and nugget=.026 for exponential, and partial sill=.018, range=1961 meters and nugget=.028 for spherical function.

$$W_{exponential}= C(h)= \begin{cases} .014e^{\frac{-||h||}{481.8}}, & \text{when } h>0 \\ .04, & \text{when } h=0 \end{cases}$$

$$W_{spherical}= C(h)= \begin{cases} .018(1-\frac{3}{2}\cdot\frac{||h||}{1961}+\frac{1}{2}\cdot(\frac{||h^3||}{1961})), & \text{when } 0<h\leq 1961 \\ 0, & \text{when } h> 1961 \\ .046, & \text{when } h=0 \end{cases}$$

Table 4: Values for Exponential and Spherical Covariance Parameters

Parameters	**Exponential**	**Spherical**
Partial sill=τ^2	.014	.018
Nugget=σ^2	0.026	.028
Range=θ	481.8	1961
Sill=$\tau^2+\sigma^2$	.04	.046
Nugget-to-sill ratio	65%	60.87%

Since the estimation of these covariance parameters was based on the parametric method, some inferences can be made. In the case of both models, autocorrelation does not seem to be very strong. This is because the range for the model with ECF is 481.8 meters, and 1961 meters for the model with SCF. In the case of spherical covariance range, covariance between two properties after 1961 meters is 0. In the case of exponential covariance range, autocorrelation tends to decay towards zero after 1445.4 meters, which is the effective range. Since the sill exists in both functions, spatial dependence does exist. Non-spatial variance in both the models was very high. That could be inferred to the fact that the data collection procedure resulted in measurement errors or that differences in sales prices of houses in closer proximity were very high. The nugget-to-sill ratio has been used primarily to quantify the issue of spatial dependency. Atkinson and Lloyd (2010) cited Wei's (2007) work where she developed three categories in determining if the spatial process has high, moderate, and low spatial dependence. Based on Wei's categorization, the nugget-to-sill ratio in both the models were in between 25% and 75%, thereby indicating a moderate spatial dependence among properties.

While the sill value represents the total variability, including the nugget variance, partial sill represents the variability that can be alluded to spatial reasons. The value for partial sill in table 4 is less than the nugget variance based on both functions, thereby a clear indication of less spatial variability but more variability in terms of non-spatial errors. When compared to the estimates in table 3 with that of the final estimates in table 4, the sill values are approximately similar. The range value has decreased and nugget variance has increased. The parametric modeling method suggested more non-spatial error and moderate spatial dependency, whereas the experimental variogram suggested more spatial error and relatively strong autocorrelation.

5.3. Model Selection

After estimating the elements of the covariance parameters, regression models using exponential and spherical covariance functions were developed. In both models, bedrooms, bathrooms, flood region, and difference in the longitude were insignificant. Likelihood ratio test (LRT) on both models were conducted to assess if the four variables could be excluded from the model. At first, a linear regression model was created excluding those four variables, and initial values for its covariance parameters were extracted from the experimental variogram of its residuals. The process followed similar procedures as discussed in section 5.1 and 5.2. The reduced models were finally created and the likelihood ratio test was performed.

The results of the LRT for both models are produced in table 5 and 6, respectively. In each of these tests, the results supported the reduced models, so the reduced models were used to make inference as well as perform regression kriging. The existence of small scale spatial variability showed the apparent need of geostatistical models, thereby OLR model was no longer considered, but the OLR model will later be discussed in the application of regression kriging techniques.

Table 5: LRT Test for the Model with ECF

Model	LogLik	Df	Chisq	Pr(>Chisq)
Full	369.794			
Reduced	368.751	-4	2.086	0.72

Table 6: LRT Test for the Model with SCF

Model	LogLik	Df	Chisq	Pr(>Chisq)
Full	363.569			
Reduced	361.924	-4	3.291	0.510

The regression coefficient output for the model that used exponential covariance function (ECF) and the model that used spherical covariance function (SCF) is given below in table 7 and table 8, respectively. In terms of numerical variables, the negative sign on the estimates of age

clearly indicated that age negatively impacts the price, whereas total area of a building, segment square feet of a property and garages seemed to positively impact the price. Compared to houses that have full basement finish, houses with less than a quarter of basement finish or more than half of the basement finish, on average, resulted in negatively impacting the price of a house. Likewise, single family dwellings seem to positively impact the prices, compared to the non-single family dwellings. Similarly, inferences for all the other variables used can be made based on the coefficient estimates in table 7 and table 8.

Table 7: Coefficient Output for the Model with ECF

names	**Estimates**	**Std.Error**	**Test Statistics**	**P**
(Intercept)	12.213	0.024	514.304	0
age	-0.101	0.014	-7.059	0
SegSqFt	0.027	0.008	3.503	0
BldgTotSF	0.221	0.01	21.706	0
basementfinishhalformore	-0.036	0.015	-2.47	0.014
basementfinishQuarterorless	-0.129	0.015	-8.805	0
garagetypeAttached	-0.082	0.018	-4.484	0
proptypeSF	0.163	0.015	10.503	0
storyheightOS	0.036	0.018	1.941	0.053
storyheightSL	-0.149	0.029	-5.227	0
storyheightTS	-0.031	0.025	-1.272	0.204
AircondNone	-0.072	0.017	-4.345	0
AircondWall	-0.072	0.019	-3.826	0
garages	0.067	0.007	10.129	0

Table 8: Coefficient Output for the Model with SCF

names1	Estimates	Std.Error	Test Statistics	P
(Intercept)	12.214	0.028	437.446	0
age	-0.087	0.016	-5.522	0
SegSqFt	0.025	0.008	3.257	0.001
BldgTotSF	0.224	0.01	21.883	0
basementfinishhalformore	-0.036	0.015	-2.453	0.014
basementfinishQuarterorless	-0.131	0.015	-8.936	0
garagetypeAttached	-0.076	0.018	-4.155	0
proptypeSF	0.163	0.015	10.651	0
storyheightOS	0.038	0.018	2.066	0.039
storyheightSL	-0.153	0.028	-5.383	0
storyheightTS	-0.035	0.025	-1.435	0.152
AircondNone	-0.072	0.017	-4.292	0
AircondWall	-0.071	0.019	-3.802	0
garages	0.068	0.007	10.144	0

5.4. Model Diagnostics

5.4.1. The Model with ECF

Figure 2 shows the histogram of the residuals for the model that used exponential covariance function. The histogram is approximately bell shaped, therefore, satisfaction of the normality of the error terms can be assumed. Figures 3 and 4 were created to assess if the model residuals have a similar pattern throughout the entire region of Fargo. Based on both plots, it was noticed that the concentration of the residuals was random in every region in Fargo. Furthermore, the plot on figure 4 was created to assess if the model either over predicted or under predicted in some regions of Fargo. When the residuals were greater than 0, they were categorized into under predicted categories, whereas residuals less than 0 were grouped into over predicted categories. Based on figure 4, it was concluded that the pattern of over and under prediction was fairly random.

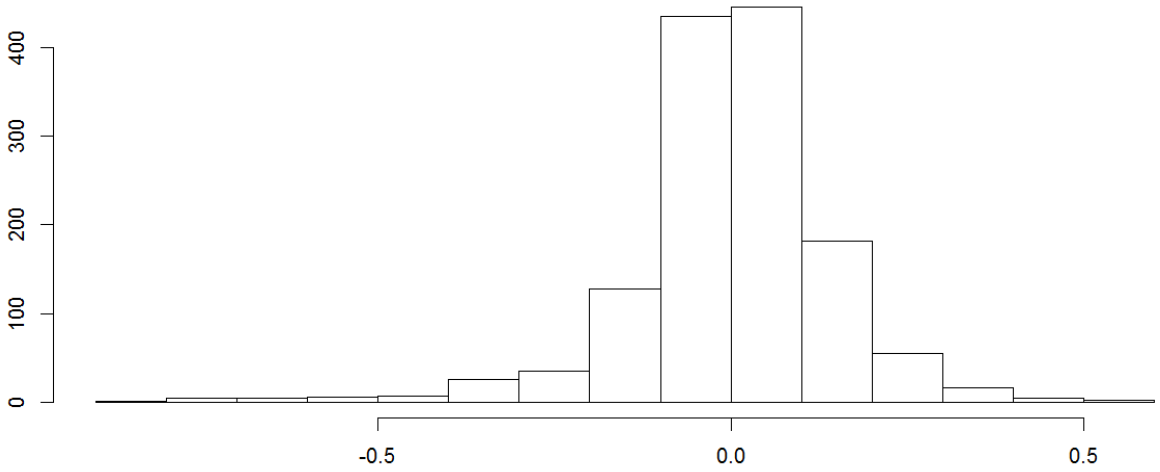

Figure 2: Histogram of the Residuals for the Model with ECF

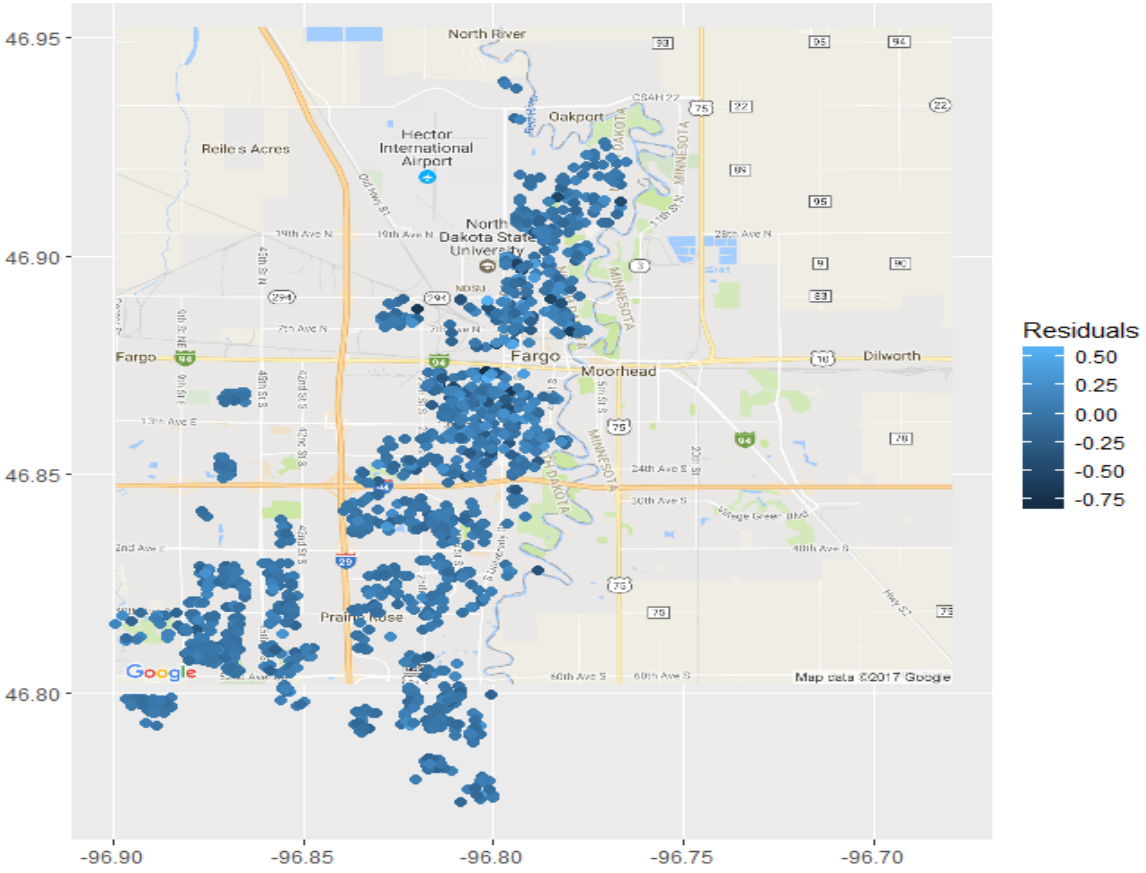

Figure 3: The Model with ECF Residuals Plot

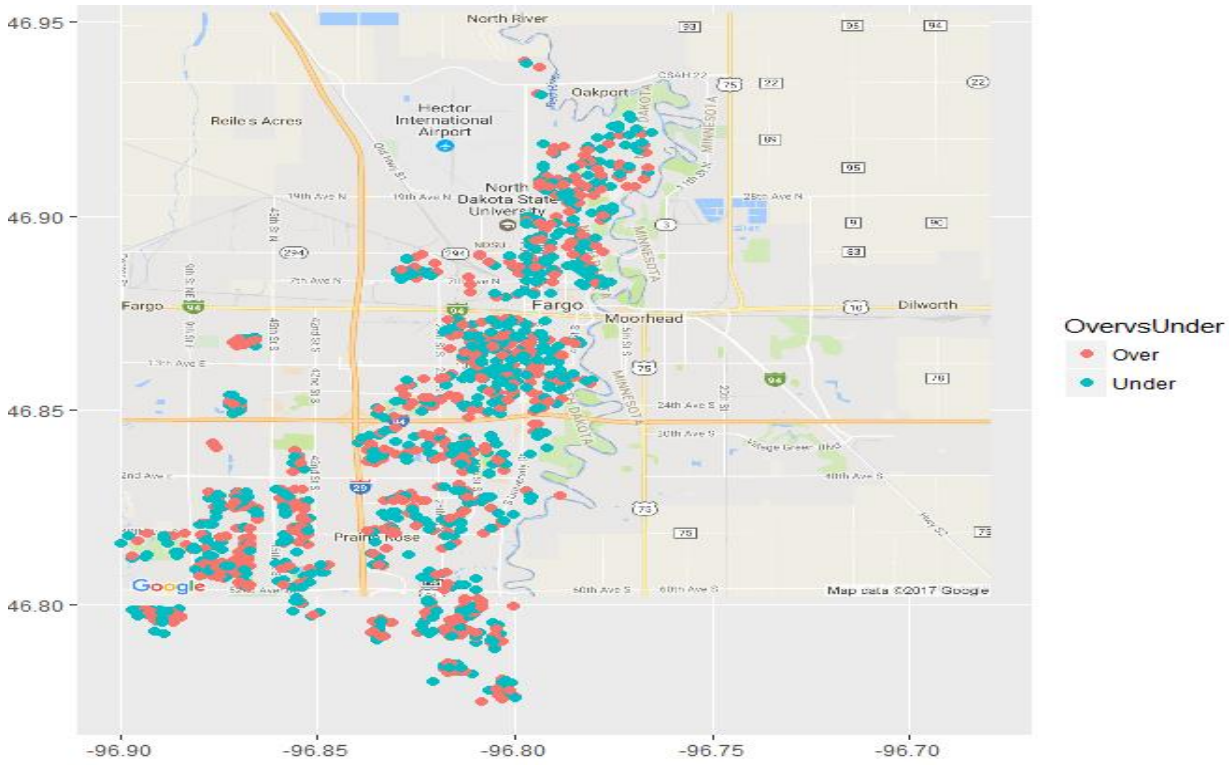

Figure 4: Over vs Under Prediction Plot using ECF

5.4.2. The Model with SCF

Similarly, figure 5 shows the histogram of the residuals for the model that was fitted using spherical covariance function. Based on the approximately bell shaped shape of the histogram, the normality of the error terms was assumed. Likewise, figures 6 and 7 were created to assess any spatial patterns of the residuals. Based on both the figures, it was concluded that the concentration of the residuals was very random in every region of Fargo. The pattern of the residuals in the geostatistical models that employed exponential and spherical covariance functions seemed to be very similar. The plot on figure 7 was generated using the over versus under prediction technique discussed in previous section

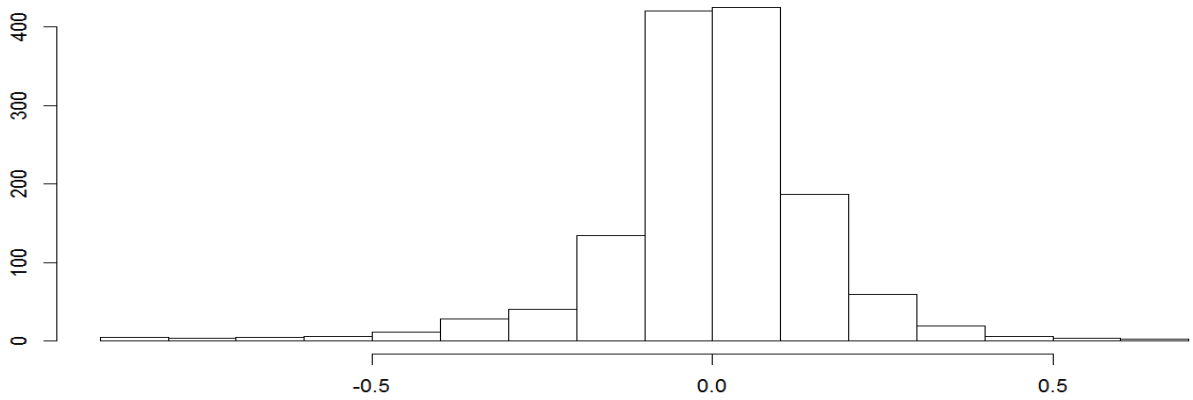

Figure 5: Histogram of the Residuals

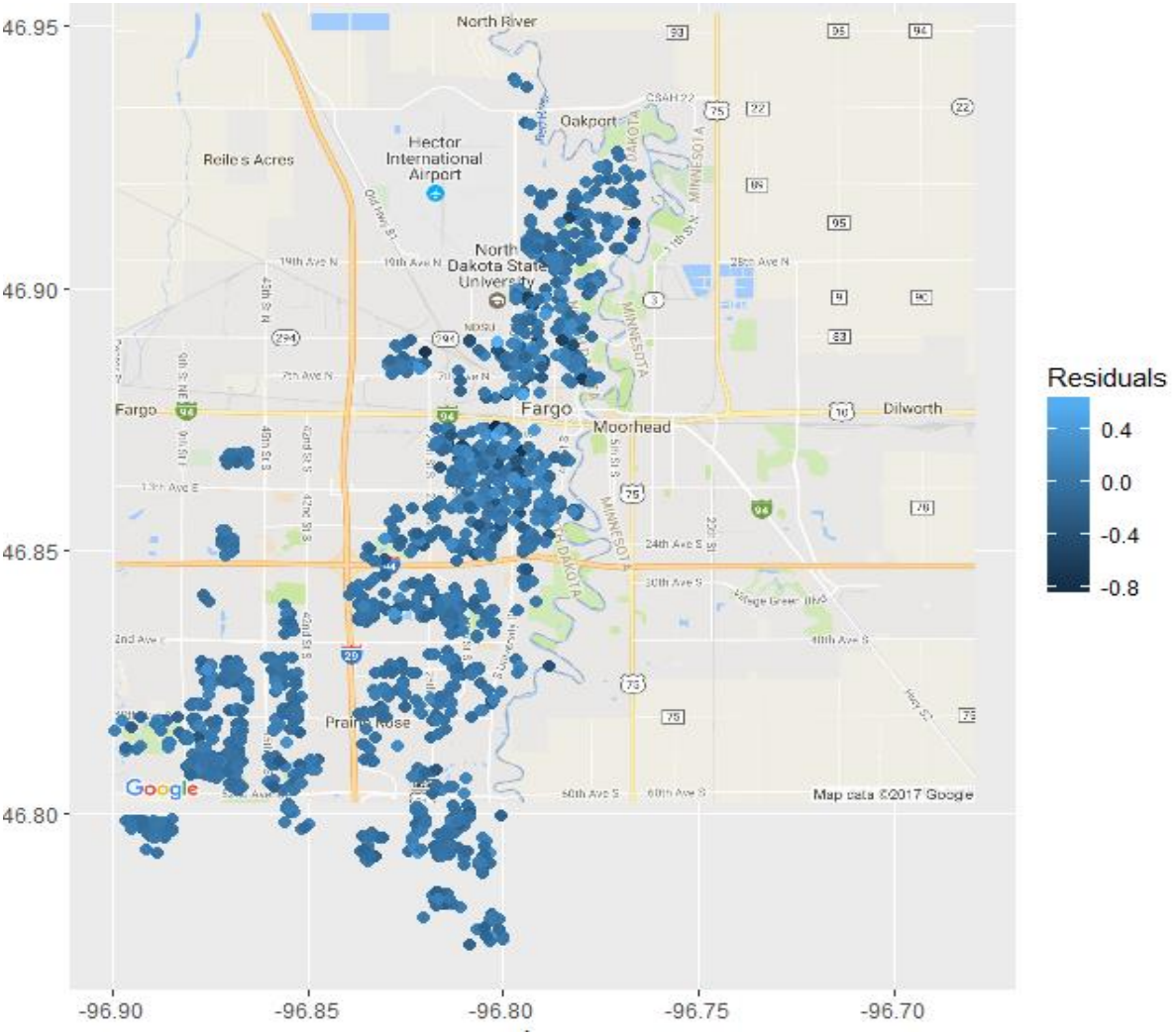

Figure 6: The Model with SCF Residuals Plot

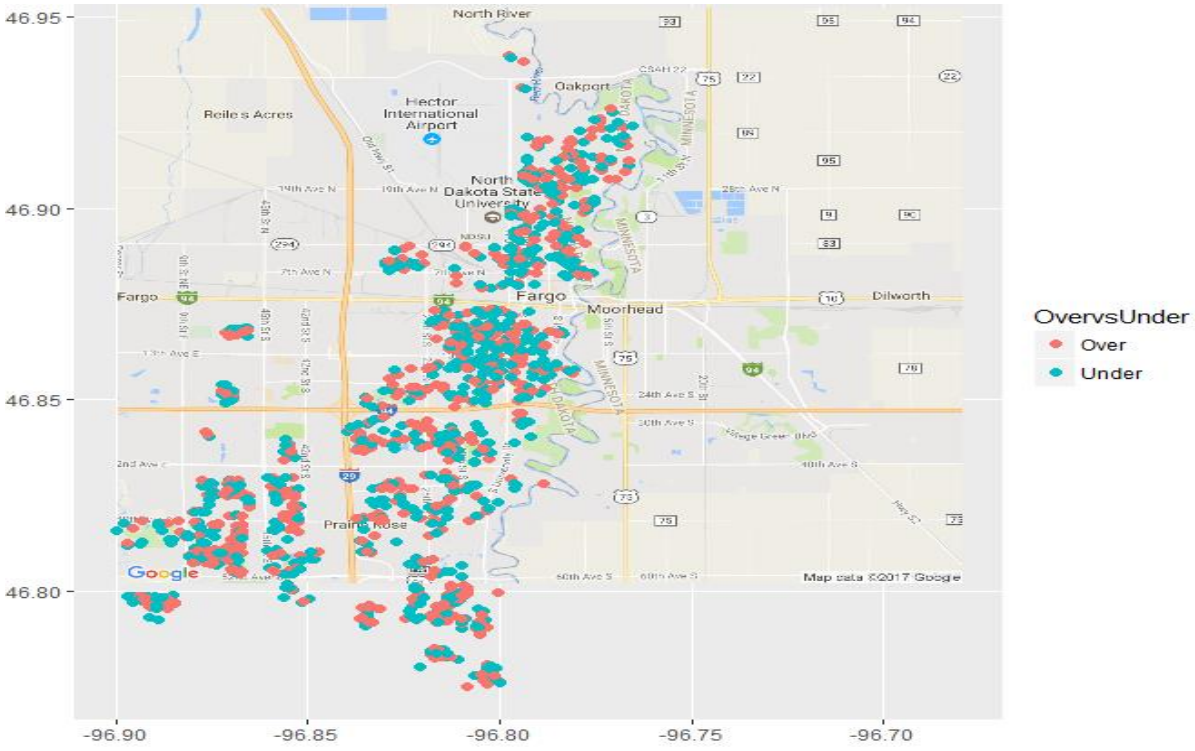

Figure 7: Over vs Under Prediction Plot using SCF

5.5. Application of Regression Kriging

As discussed in section 4.6, regression kriging technique was applied to predict new data values at new locations. The prediction was performed on a dataset with 341 observations. Then the covariance vector of residuals at new prediction locations for exponential and spherical functions were developed as:

$$g_0 = \{.014e^{\frac{-||h||}{481.8}}\}$$

$$g_0 = \left\{.018(1 - \frac{3}{2} \cdot \frac{||h||}{1961} + \frac{1}{2} \cdot \frac{||h^3||}{1961}) \right\}$$

, where ||h|| is the Euclidean distance between the original data locations and new location. Based on equation (4.11), new data values were predicted at new locations. The map of the predicted values is given in figures 8 and 9, respectively for the values predicted using ECF and SCF. Three important conclusions were derived from the map:

- The South Fargo region was predicted to have houses with high property values around the boundary of South University Drive, Interstate 94 and towards the border with West Fargo. Likewise, the North Fargo region up from 19th Ave North was found to be the second major region with a large concentration of high property values.
- The concentration of high property values in those two main regions may be due to the fact that much of the development activities in Fargo is in its southern side, whereas later developments in residential activities in the North Fargo side may be attributed to the larger concentration of high residential properties.
- As the coefficients' output showed the impact of independent covariates on the prices, it could be that the recently built houses have higher values because of being recently built along with other core structural advantages such as increase in building size, which older houses may lack.

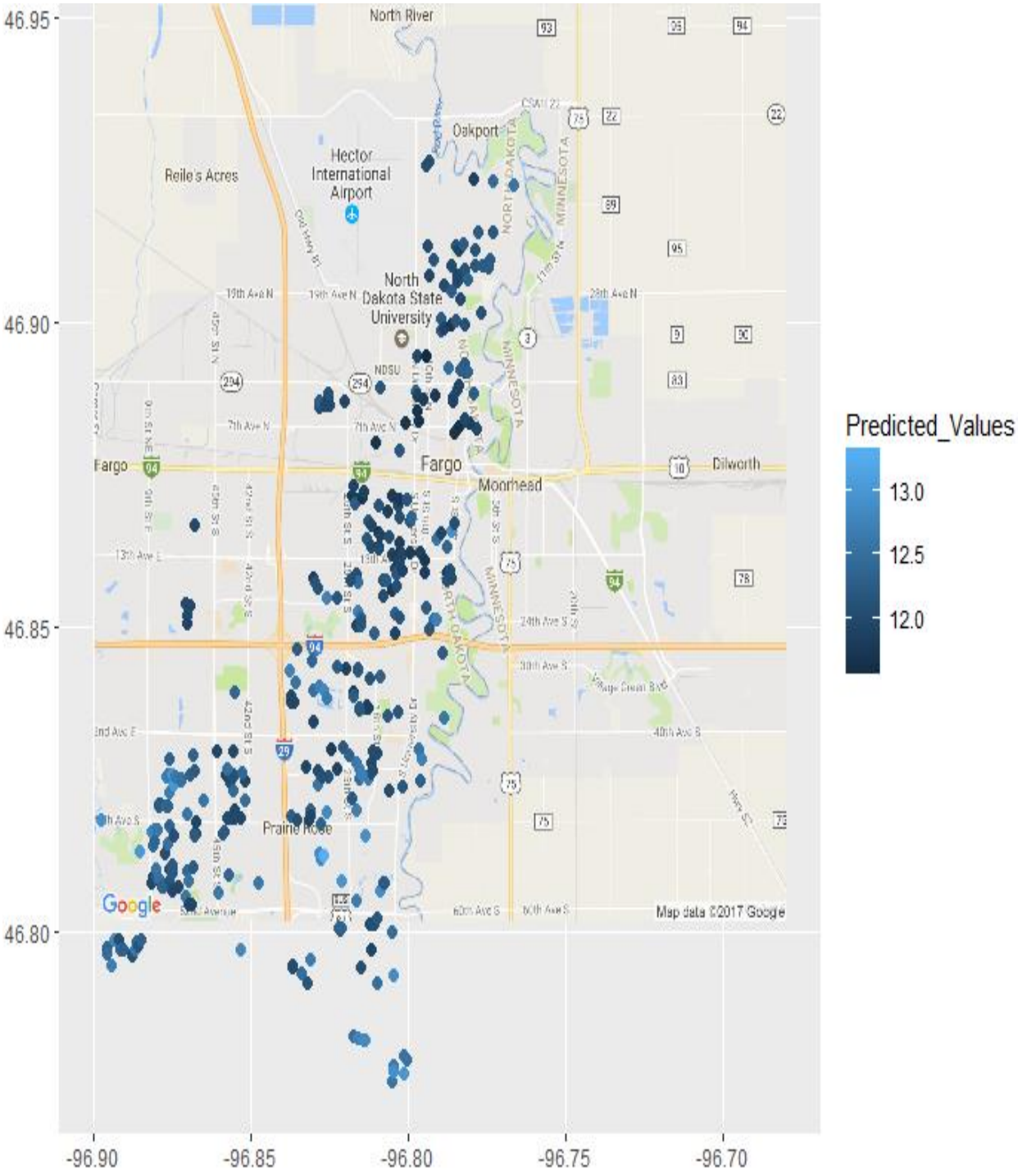

Figure 8: Regression Kriging Estimated Values Plot using ECF

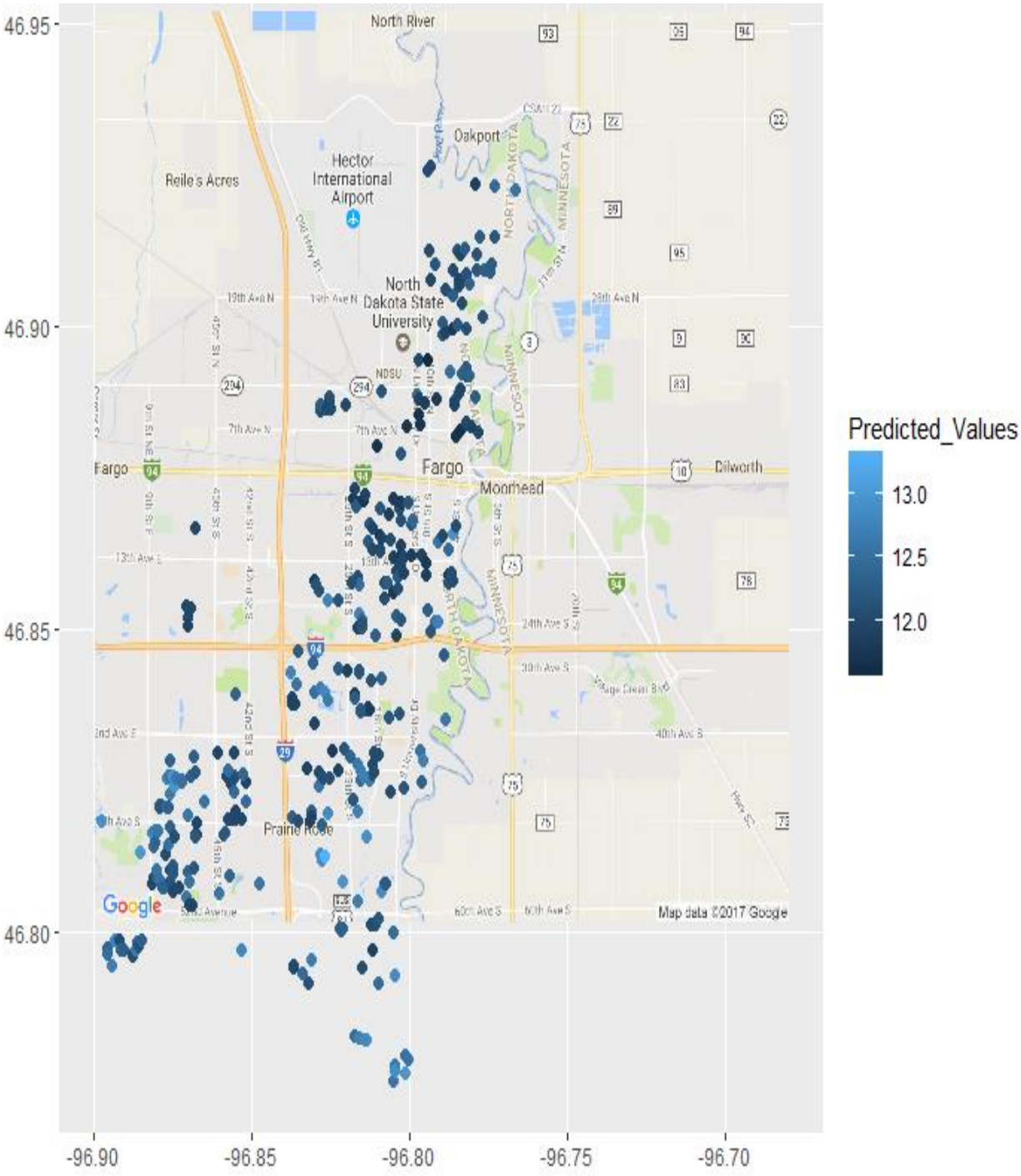

Figure 9: Regression Kriging Estimated Values Plot using SCF

After predicting new data values, confidence interval on the new predicted values were calculated to assess how well the actual values would fall in between the 95% confidence limit of new predicted values. Thus, the OLR model and geostatistical models with ECF and SCF were used to predict prices of houses at new locations. A sample of ten new houses with their predicted values, along with the standard errors is given in table 9. Based on the table below, the standard errors for new predicted values using the OLR model was much smaller than the geostatistical models with ECF and SCF. The standard errors for the new predicted values based on the geostatistical techniques seemed to be very similar.

Table 9: A Sample of 10 new Houses with Predicted Prices

Observed log price	**OLR**	**ECF**	**SCF**
13.305	13.049	12.945	12.925
	(.017)	(.189)	(.191)
12.824	12.472	12.384	12.375
	(.016)	(.178)	(.179)
12.656	12.577	12.481	12.473
	(.016)	(.181)	(.181)
12.814	12.821	12.713	17.707
	(.016)	(.188)	(.187)
12.801	12.951	12.808	12.799
	(.024)	(.176)	(.178)
12.230	12.150	12.067	12.057
	(.014)	(.176)	(.178)
12.407	12.477	12.381	12.373
	(.016)	(.172)	(.175)
12.268	12.424	12.310	12.302
	(.021)	(.171)	(.174)
12.186	12.285	12.162	12.156
	(.023)	(.175)	(.176)
12.219	12.162	12.069	12.064
	(.015)	(.177)	(.178)

Furthermore, if the actual prices were within the 95% confidence interval of the predicted values, then they were categorized as being “inside”, while those that were less than the lower limit of the confidence interval were categorized as “less” and those that were above the upper

limit of the confidence interval were categorized as “over”. Based on the output in table 10, the geostatistical models included approximately around 90% of the actual values inside the 95% confidence interval of the predicted values, whereas the OLR model was only able to include around 18% of the actual values within the 95% confidence interval of the values it predicted.

Table 10: Comparison of RK and OLR C.I. Prediction

Groups	OLR	ECF	SCF
Inside	61	306	302
Less	158	27	30
Over	122	8	9

However, it was also noticed that the average width of the OLR model predicted values confidence interval was much smaller than the average width of the geostatistical models predicted values confidence interval. However, emphasis should be given to the fact that, unlike the case of OLR prediction, the geostatistical modelling prediction error relies on the deterministic part as well as the residual kriging component. The output on table 11 shows the average width of the predicted values confidence limit.

Table 11: Average Width

OLR	ECF	SCF
.074	.707	.708

The plots on figures 9, 10 and 11 show the locations where confidence interval on the predicted values were calculated. The plots on figures 9 and 10 relate to the prediction of new values based on regression kriging technique using exponential and spherical covariance functions, respectively. Likewise, the plot on figure 11 relates to the prediction of new values using OLR model. Based on figures 9 and 10, it was concluded that the geostatistical models included most of the actual values in different regions of Fargo in the predicted values confidence interval.

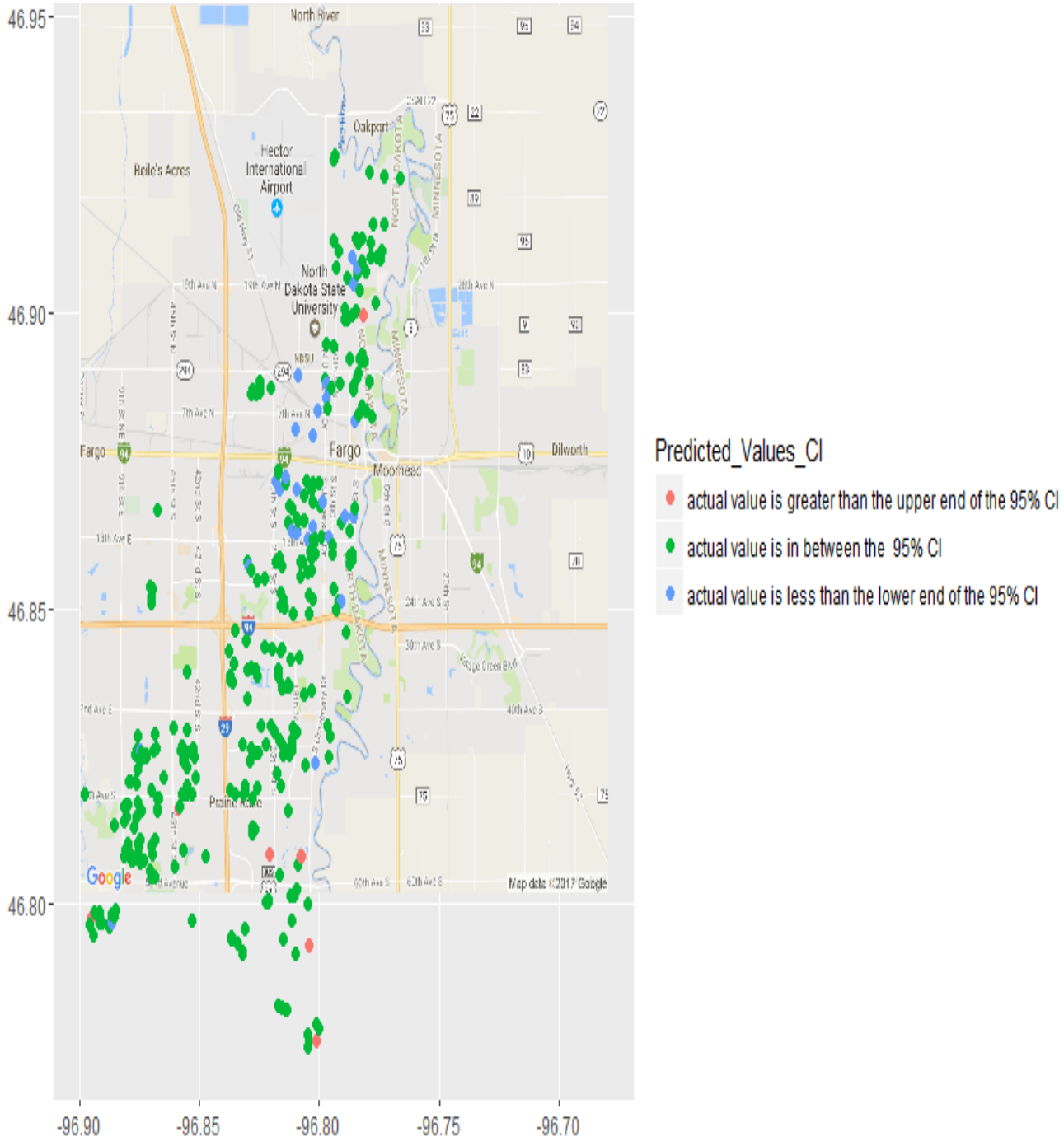

Figure 10: Confidence Interval Plot of Predicted Values using ECF

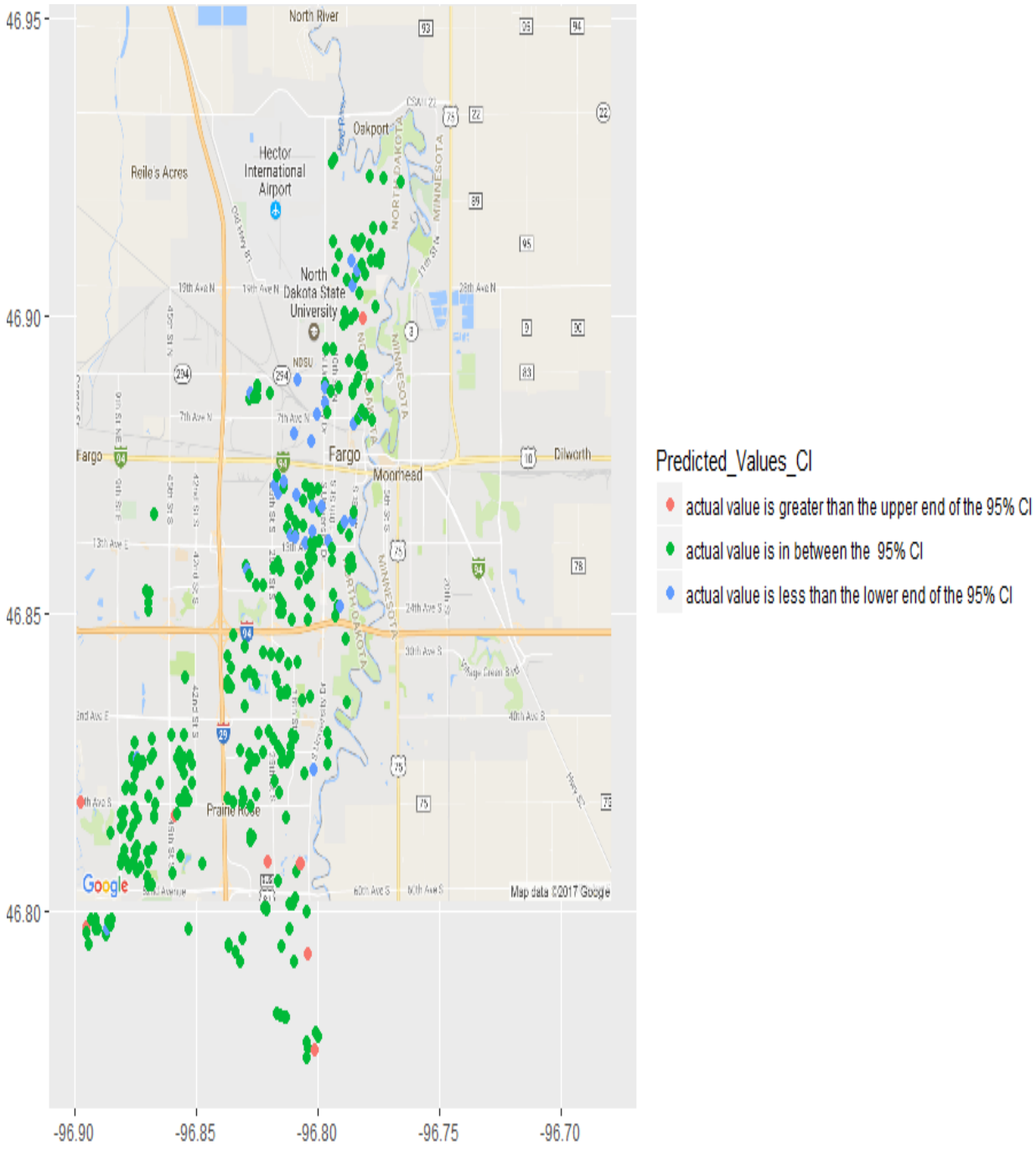

Figure 11: Confidence Interval Plot of Predicted Values using SCF

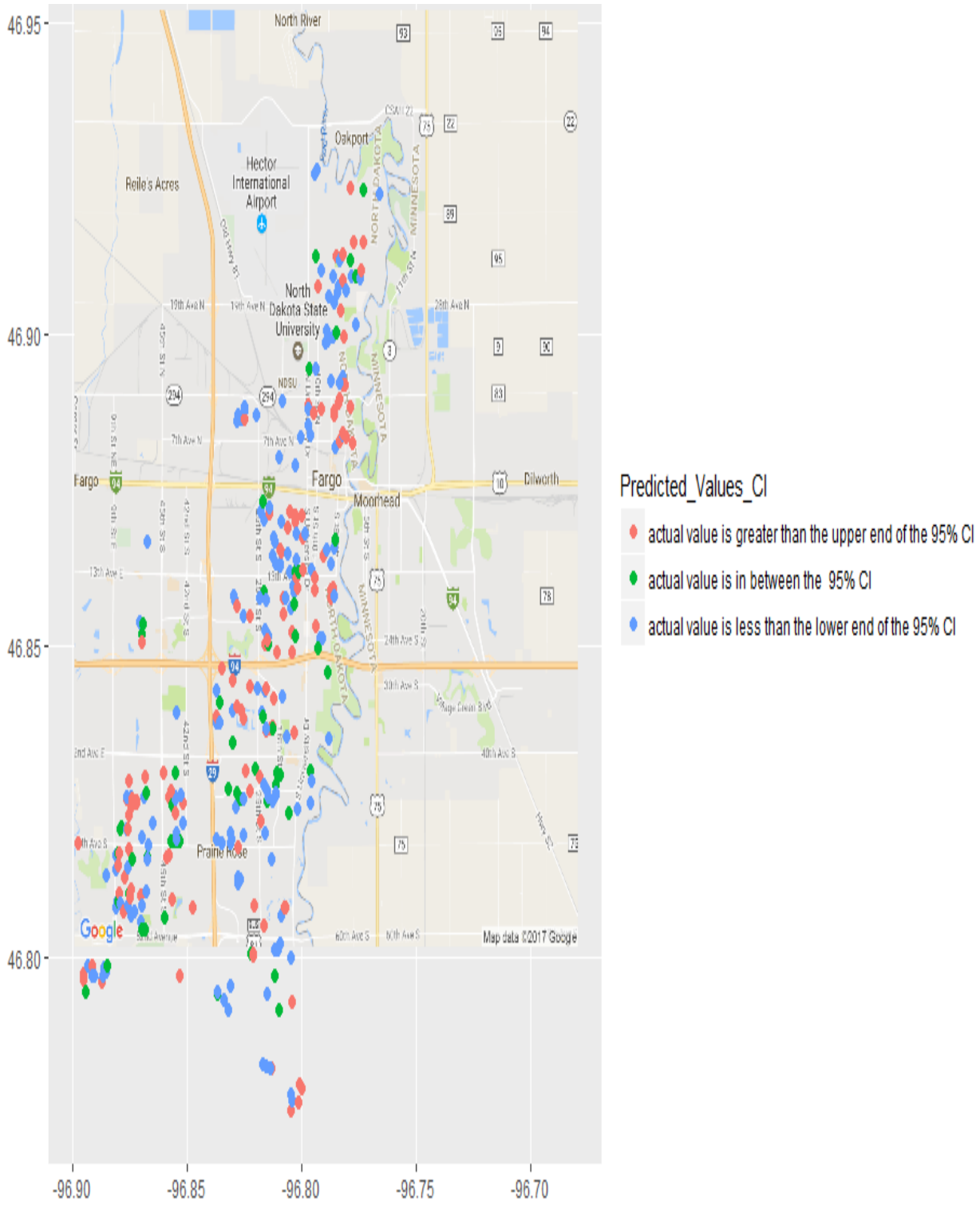

Figure 12: Confidence Interval Plot of Predicted Values using OLR

6. CONCLUSION

In this thesis, approach to modeling housing prices began with ordinary linear regression model. However, based on the parametric estimation of the covariance parameters, the existence of small-scale spatial variability was detected. Thus, the spatial covariance parameters were included in the modeling approach by using them as a function of distance. Age, area of the property, garage type, property type, style of the house, air conditioner type, status of basement finish and number of parking spots inside a garage were found to be statistically significant factors in determining housing prices in Fargo. Unlike past research in the field of real estate appraisal, this thesis did not find number of bathrooms, number of bedrooms and flooding history to be statistically significant. The non-spatial error variance amounted to most of the total variance, but using the small-scale spatial error in the modeling approach performed better when predictions on new locations were made. While moderate spatial dependency was detected, the application of regression kriging technique was implemented to perform new predictions.

While the inclusion of geostatistical technique in the field of statistical research has been growing, improvement of some of these techniques is also necessary. Future work in the field of geostatistical research could include proper variogram binning technique. Furthermore, the government appraisal technique, if not already, should start incorporating the spatial elements into their appraisal methodologies. Such inclusion, as the findings of this thesis show, may yield better predictive modeling.

REFERENCES

Abraham, B., & Ledolter, J. (2006). *Introduction to regression modeling*. Belmont, CA: Thomson Brooks/Cole.

Atkinson, P. M., & Lloyd, C. D. (2010). *GeoENV VII - geostatistics for environmental applications: proceedings of the seventh European Conference on Geostatistics for Environmental Applications*. Dordrecht: Springer.

Bishop, A. (2016). Valley News Live. Will increased property values change the housing market in the FM area? *Fargo Forum*. Retrieved from http://www.valleynewslive.com/home/headlines/Will-increased-property-values-change-the-housing-market-in-the-FM-area----375321271.html.

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2007). Spatial dependence, housing submarkets, and house price prediction. *The Journal of Real Estate Finance and Economics*, 35, 143-160.

City of Fargo. (n.d.). What is mass appraisal. Retrieved from https://www.cityoffargo.com/CityInfo/Departments/Assessor/AssessmentProcess/MassAppraisal.

Diggle, P., & Ribeiro, P. J. (2010). *Model-based geostatistics*. New York, NY: Springer New York.

Dubin, R. A. (1998). Predicting house prices using multiple listings data. *Journal of Real EstateFinance and Economics* , 17:1, 35-59.

Gelfand, A. E., Ecker, M. D., Knight, J. R., & Sirmans, C. F. (2004). The dynamics of location in home price. *Journal of Real Estate Finance and Economics*, 29:2, 149-166.

He, C., Wang, Z., Guo, H., Sheng, H., Zhou, R., & Yang, Y. (2010). Driving forces analysis for residential housing price in Beijing. *Procedia Environmental Sciences*, 2, 925-936.

Hengl, T., Heuvelink, G. B., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computers & Geosciences*, 33, 1301-1315.

Kahle, D., & Wickham, H. ggmap: spatial visualization with ggplot2. *The R Journal*, 5:1, 144-161.

Kim, S. (2015). The estimation of the variogram in geostatistical data with outliers. Retrieved March 02, 2017, from http://ousar.lib.okayama-u.ac.jp/files/public/5/53442/20160528121035248979/K0005160_fulltext.pd.

Kitanidis, P. K. (2003). *Introduction to geostatistics: Applications to hydrogeology*. Cambridge: Cambridge Univ. Press.

Koramaz, T. K., & Dokmeci, V. (2012). Spatial determinants of housing price values in Istanbul. *European Planning Studies*, 20:7, 1221-1237.

Monson, M. (2009). Valuation using hedonic pricing models. *Cornell Real Estate Review*, 7, 62-73.

Olmo, J. C. (2007). Prediction of housing location price by a multivariate spatial method: Cokriging. *JRER* , 29:1,91-113.