Modeling church services supply and performance, using geographically weighted regression

Xin HE

October, 2009

Master’s Thesis in Geomatics
# Table of Content

Abstract............................................................................................................................... 2

1 Introduction....................................................................................................................... 3
  1.1 Model development ................................................................................................. 3
  1.2 Earlier studies and gaps ......................................................................................... 4
  1.3 Objective of this study .......................................................................................... 6

2 Statistical concepts ....................................................................................................... 7
  2.1 Possible solutions ................................................................................................. 7
  2.2 Using the lagged dependent variable ...................................................................... 8
  2.3 First-differencing estimation for two-period panel data ........................................ 10
  2.4 Geographically weighted regression ..................................................................... 11
     2.4.1 Local model and its estimation ................................................................ 11
     2.4.2 Weights matrices ....................................................................................... 14
     2.4.3 Model calibration ....................................................................................... 15
     2.4.4 Significant tests for spatial variations ....................................................... 15
  2.5 Some statistical tests ............................................................................................ 18
     2.5.1 t-test ............................................................................................................... 18
     2.5.2 F-test ............................................................................................................ 19
     2.5.3 White Test .................................................................................................... 20
     2.5.4 RESET test .................................................................................................. 20

3 Materials and methods ............................................................................................... 22
  3.1 Materials and data processing ............................................................................. 22
  3.2 Models .................................................................................................................. 23
  3.3 Modeling strategies comparison and GWR estimation ....................................... 24

4 Results and analysis ................................................................................................... 26
  4.1 Performance of the non-geographical regression models ................................... 26
  4.2 Selection of variables .......................................................................................... 30
  4.3 GWR estimations ................................................................................................. 32
  4.4 Comparison between OLS and GWR ................................................................. 37
  4.5 Model interpretation ............................................................................................ 40

5 Discussions ............................................................................................................... 42

6 Conclusions ............................................................................................................... 44

Acknowledge .................................................................................................................. 45

Reference ....................................................................................................................... 46
Abstract

The objective of this study is to develop a multiple linear regression model that measures the relationship between the church services supply and the attendance to the services in the Uppsala diocese, Church of Sweden. By reviewing previous models and examining the nature of data available, two research questions were introduced, namely, the problem of omitted variables and the problem of spatial autocorrelation. For the first question, two methods were compared, namely, the Y-lag method and the first-differenced equation. Statistical tests then showed that the latter was more preferable for this study. For the second question, geographically weighted regression was used to examine the spatial variations in relationships estimated by above modeling strategies. However, no significant spatial variation was found for them. In conclusion, by using the ordinary least square estimation for the first-differenced equation the most suitable regression model was obtained. The data showed no need to consider the issue of spatial non-stationarity.
1 Introduction

The Church Ordinance states that (Svenska Kyrkan, 1999, cited by Lind, 2003), “The church of Sweden appears locally as parish. This is the primary unit of the church.” The Uppsala diocese has about 169 parishes, which are different in geographic size, population, and socio-economic conditions. To manage these diverse parishes is the main task of the Uppsala diocese. It demands a clear view on how much the church service supply at each parish can affect their performances, i.e. the attendance to the services, by considering the geographic, demographic, and socio-economic factors.

Some relevant studies have been performed in Sweden. However, few of them are carried out at the parish level. Therefore, there is probably no systematic model describing that relationship for local parishes. It could be an obstacle that prevents the parishes and therefore the dioceses to achieve a more efficient way of management and to get a better performance. So models that describe the relationship between the church service supply and their performance at the parish level are demanded. This study attempts to develop one of such model for the parishes of Uppsala diocese. This type of models may be used to identify under-performing parishes, which may result in additional support actions. In addition, the models may also be used in evaluating structure changes in the organization.

1.1 Model development

To develop a model for the parishes of Uppsala diocese, several methodologies could be used. A common way is to develop an analytical model from a set of basic assumptions, as the rational assumption in economics (Becker, 1976), and then validate and verify this model by the statistical test when the data are available. Another approach is to directly carry out the statistical modeling by computerized techniques such as data-mining.

This study employs a way of modeling that is more similar to the first methodology. However, because of the limitations in research time and resources, this study doesn’t start from the very beginning of an analytical analysis that studies the mechanism behind the parish operations. Instead, the regression models from studies of Hamberg and Pettersson (1994) and Berggren (1997) are used as the starting point. Analysis is then performed to find out the differences between those studies and this one. Efforts are made to search ways to make up for these differences, and finally to develop the regression model for the Uppsala diocese.
1.2 Earlier studies and gaps

The structure of the religious market has been studied since the mid of 1980s. Several studies before the work of Hamberg and Pettersson (1994) indicate that in the religious markets, as in other kinds of markets, more competition brings more religious “goods” or services. However this conclusion is drawn in terms of American data. It therefore motivated Hamberg and Pettersson to test it with Swedish data. Their work is carried out with help of the Swedish Church statistics and the Swedish census data and finally makes the American theory be accepted in Sweden.

On the other hand, Berggren (1997) develops a theory that says the Christian religious involvement in Sweden has negative effects on some Swedish social problems as the divorce, abortion, non-payment of the bill, and children-born out of the marriage. This theory is developed in terms of the classical assumption of rational choice. Multiple linear regressions (MLR) with a cross-sectional data set in time are employed to test and finally to accept the theory.

However, those researches are different to this study in at least three aspects. First, the aims of the modeling are different. Hamberg and Pettersson (1994) examine the structure of the Swedish religious market, whereas Berggren (1997) provides an economic analysis for effects of religious involvement on some social problems. Second, they are developed at the municipal level which is less detailed than the parish level used in this study. Third, since the statistical units are different, many data that are accessible at the municipal level are unavailable at the parish level. Therefore many factors that are considered in these previous models can not be included in this study.

These differences suggest several problems this study might suffer from. The first and the most crucial one is the problem of omitted variables (Woodbridge, 2002). It happens if a factor is omitted from a regression model and this omitted factor both influences the dependent variable and is correlated to some other independent variables of this regression model. In such a case, the estimation of the model parameters will be biased.

Unlike the model of Hamberg and Pettersson (1994), this study lacks measurements for some socioeconomic and demographic factors such as the ratio of the population with annual salary higher than 200,000 SEK, or the percentage of the population living outside
densely populated areas, etc. These factors, although no proof is found, very likely have influences on the church services supply in each parish. If that is the case, the estimation obtained by the method of ordinary least square (OLS) would be biased.

Another issue is the modifiable areal unit problem (Openshaw, 1984, Fotheringham et al. 1995, 2002). The modifiable areal unit problem (MAUP) means that different results can be obtained because the analyses are carried out at different levels or different zoning systems at a given spatial scale. It might cause the results from previous models incomparable to this study. Geographically weight regression (GWR) (Brunsdon et al., 1996; Fotheringham et al., 1997 a,b, 1998, 2002; Leung et al., 2000a,b; Huang and Leung, 2002) could be used to reduce the effects of MAUP. However none of the above studies employ this approach. So even if the GWR is applied here, the results of this study may still be incompatible with the previous studies.

Besides, the issue of spatial non-stationarity isn’t covered by previous Swedish studies. But it is a main concern of this study. The spatial non-stationarity means that the relationship of concern may vary across the studying space (Fotheringham et al., 2002). For a specific model, the results of regression are then different at different places.

Spatial non-stationarity has bothered socioeconomic scientists for many years and whether it really exists is still in debate. Fotheringham et al. (1997a,b, 2002) review these debates and summarize reasons why spatial non-stationarity is of interest. The first is the variation of random sampling. Suppose data are collected at different places and models are developed for each of them. Even if the actual relationship is spatial stationary and the model is perfectly specified, for the differences in the randomly sampled data, the estimations of the parameters are unlikely the same. The second explanation could be, if possible, that some relationships are naturally different in space. Because the culture and tradition are diverse, the patterns of human behavior are probably different at different place. It could therefore cause the socioeconomic relationships to become heterogeneous over space. The third possible cause is the problem of omitted variables or of model misspecification. If this is the case, the testing for spatial non-stationarity could then be used as a diagnose tool that indicates the problems with the model.
1.3 **Objective of this study**

This study tries to develop a suitable regression model to describe the relationship between the church services supply and their performance at the parish level. In this study, the term “suitable” has at least three meanings. First, the parameters of the models should be interpretable. Variables should be carefully chosen so that their parameters could have practical meanings. Second, as it is required in other statistical modeling work, the estimators of the model parameter should meet the requirements of unbiased or consistency. These requirements demand that the expected value of the estimated parameter should be equal to its actual value in any or at least in large sample size. Otherwise, the model would lose most of its explanatory ability and have less practical value. Third, an advanced statistical tool, such as GWR, could produce a higher degree of explanation which may be insignificant in the sense either of statistics or of economics. In such a case, the simpler technique that provides the weaker but similar explanation should be adopted. It is because that the model or its result is the end users’ wants. Simple model in most instances is easy for interpretation and application. So there is no need to use the advanced model just because it looks complex.

To achieve this goal, the rest part of this study is to

- Compare different types of regression models with respect to church performance;
- Estimate parameters of the regression model for the Uppsala diocese.
2 Statistical concepts

2.1 Possible solutions

The main task of this study is to find ways for solving or releasing the problems of omitted variables and spatial non-stationarity. For the omitted variables problem, the most direct and efficient treatment is to include these omitted variables in the regression model. However it is usually impossible due to limitations in data collection. Luckily, statisticians and econometrician have developed series of statistic techniques to eliminate, or at least, to greatly reduce the biases of parameter estimations when some important variable are known excluded. These statistic techniques commonly include (Woodbridge, 2002):

- Using proxy variables for the omitted variables;
- First-differenced (FD) /fixed-effects (FE) /random-effects (RE) estimations for the panel dataset;
- Instrument variables (IV) and two-stage least square (2SLS);
- Simultaneous equations model;
- Etc…

Including a lag in time of dependent variable (Y-lag) to the model is one form of the proxy variables approach that deals with the problem of omitted variables (Woodbridge, 2002). If the omitted variables are known, there is a possibility that these variables can be replaced by proxy variables. However, the requirements for the choices of proxy variables are strict. Usually, it is difficult to obtain the proper proxy variables. To cope with this problem, Woodbridge (2002) suggests using the lag in time of dependent variable as the proxy variable. Including the Y-lag could control some “historical” or “tradition” factors which influences the current Y but are hard to be measured in other ways. Obviously, this idea is not perfect, but it could bring a “better” result to the estimation of model parameters.

For the analysis of panel dataset, the method of FD estimation was selected. A panel dataset here is a dataset collected for the same statistical units at different times. After the data preprocessing, panel data for 70 pastorats (using the definition of pastorat in 2006) were available from the year 2004 to 2006. Besides, in order to keep larger variations in the value of variables, this study only used data of the years 2004 and 2006 for the analysis. Woodbridge (2002) argues that for the data that are not obtained from the
“really” performed random sampling in a large population but just recorded annually in each of the statistic zone, the FD estimation is more suitable than the RE estimation; and for two periods panel dataset, as what was prepared in this study, the FD and FE estimations are identical. But the operation and interpretation of the FD estimation is simpler than the FE estimation. Therefore, this study employed the FD estimation.

To apply IV and 2SLS, or to develop the simultaneous equations model, the discussion about the mechanism behind church and parish operations must be carried out. However, it was not the goal of this study and therefore the use of these methods had not been tested.

For testing and measuring the potential spatial not-stationarity, the geographically weighted regression (GWR) was applied in this study. Although there are many other types of spatial analysis or techniques that can also be used for this purpose, the GWR is demonstrated to be simpler and more efficient in doing this work (Fotheringham et al., 2002). It allows the estimation of model parameter to be different at different regression points, and therefore allows the modeled relationship to vary across the space. Based on these estimations, the spatial non-stationarity can then be both tested and measured (Fotheringham et al., 2002; Leung et al., 2000a,b; Huang and Leung, 2002).

The rest part of this section aims to describe the statistical techniques employed in this study. It includes the lag of Y in time (Y-lag), the first-differencing (FD) estimation and the geographically weighted regression (GWR). Some statistical tests, as \( t \)-test, \( F \)-test, White-test, and RESET-test are introduced as well.

### 2.2 Using the lagged dependent variable

Including the lag of Y in time to the regression model is a common approach to the problem of omitted variables. (Woodbridge, 2002).

Assume we have a regression model,

\[
Y_i = \alpha_0 + \sum_{k=1}^{p} \alpha_k x_{ik} + \theta_i, \quad i = 1, \ldots, n
\]
the Ordinal Least Squares (OLS) estimation is

$$\hat{\alpha} = \left(X^T X\right)^{-1} X^T Y$$

With $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$, $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$, and $\hat{\alpha} = \begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_p \end{bmatrix}$. 

(2)

where $\hat{\alpha}$ is the estimation for model parameters, $\alpha^T = (\alpha_0 \ \alpha_1 \ \cdots \ \alpha_p)$.

If this model suffers the problem of the omitted variables, the estimation of $\alpha$ is biased, which means that

$$E(\hat{\alpha}) \neq \alpha.$$ 

(3)

In case we have data from different points in time. A possible solution for this problem is to include the lag of $Y$ in time to the regression model.

$$Y_i = \alpha_0 + \sum_{k=1}^{p} \alpha_k x_{ik} + Y_{-lag} + \theta_i, \ i = 1, \ldots, n,$$

(4)

where the $Y_{-lag}$ denotes the lag of $Y$ in time. Since including it could control some “historical” or “tradition” factors that influences the current $Y$ but are hard to be measured in other ways, the Y-lag serves as a proxy variable for these missed variables. In many cases, it could bring a “better” estimation of the model parameters.

For example a study intends to measure the contribution of the education background to the personal annual income. Only a dataset recording the personal annual income, education background, and length of work experience is available for this study. The linear regression model could be built directly by relating the annual income to the education background and length of work experience. But it could ignore some other considerable factors from the regression model. A possible factor is the type of work,
which influences a person’s annual income and which at the same time relates to his/her education background and work experience. If this is the case, according to the Gaussian-Markov assumptions for the linear regression model, the estimation for the model parameter will be biased (Woodbridge, 2002).

Including the record of personal annual income from an earlier year (the past year or several years ago) to the regression model could release this problem to some extend. Many unobserved factors, including the type of work, contribute both to the present and past personal annual income. Therefore, including the past annual income could provide a better control to these unobserved factors, which further results in a “better” estimation of the model parameters.

2.3 First-differencing estimation for two-period panel data

The FD is another method for dealing with the problem of omitted variables. This method uses the data from at least two years. The idea behind FD is different to that of Y-lag. The FD, using the panel data, develops two cross-section (of time) models for describing the relationships of concern between two years. More importantly, it considers that the errors of the regression model consist of two parts: one is constant in time and another is varying in time. A simple example that has only one independent variable can be written as,

$$Y_{it} = \alpha_0 + \delta_0 d_{2t} + \alpha_1 x_{it} + \eta_i + \mu_{it}, \quad t = 1, 2$$

(5)

where $i$ denotes the $i$th parish or observation, $t$ denotes the time periods, and $t = 1$ means the previous year. $d_{2t}$ is a dummy variable that equals to one when $t = 2$ and equals to zero when $t = 1$. Its coefficient $\delta_0$ measures the changes of $Y$ from the year 1 to 2. The variable $\eta_i$ represents the time-invariant unobserved factors, whereas the $u_{it}$ captures all time-varying unobserved factors. The $x_{it}$ is the independent variable and $\alpha_0$ and $\alpha_1$ are the parameters for this regression model.

One benefit of applying the FD estimation is that it allows the time-invariant error, $\eta_i$, to be correlated with the explanatory variables, say $a_1$ in the equation (Equation 1) (Woodbridge, 2002). Usually, the unobserved, time-invariant factors include geographical features, historical or traditional features, etc. In addition, some other
factors, for example some demographic factors as the ratios of each age group to the population, varies but the magnitudes of variation are relatively small. In practice, if the difference of the time periods is comparatively small, say two to five years, this kind of variables could be treated as fixed in time. For the single cross-section model, doing this would definitely cause the parameter estimation for the model biased, if the \( \eta_i \) has correlations with independent variables. However, during the FD estimation, the \( \eta_i \) could be excluded through subtracting the model for the year 2 by the model for the year 1, that is,

\[
(Y_{i2} - Y_{i1}) = \delta_0 + \alpha_1(x_{i2} - x_{i1}) + (\mu_2 - \mu_1),
\]

or

\[
\Delta Y_i = \delta_0 + \alpha_1 \Delta x_i + \Delta \mu_i,
\]

where the “\( \Delta \)” denotes the changes from the time period 1 to time period 2. Obvious, the coefficients in models (Equations 6 and 7) have an explanation that they measure the effects of changes of independent variables on the change of the dependent variable. However, they also share the same meanings of coefficients in the model (Equation 1).

If the \( \Delta u_i \) is uncorrelated to the \( \Delta Y_i \), then the unbiased estimation for the parameter of \( \Delta x_i \) could be made. It actually reduces the possibility of the problem of omitted variables, and the unbiased estimation could be obtained by controlling less factors. However, this requirement means the lag of Y in time can’t be included as the explanatory variables (Woodbridge, 2002). It therefore rules out the possibility of combining the uses of Y-lag and FD together. As a result, a comparison between the Y-lag and FD should be made.

### 2.4 Geographically weighted regression

#### 2.4.1 Local model and its estimation

As it states by Fotheringham et al. (1997a, b, 1998, 2002), Leung et al. (2000a, b) and Huang and Leung (2002), GWR is an efficient tool for calculating the local estimations
for the model parameters and for testing the spatial non-stationarity. It provides a method
to estimate the model parameter for each of the regression point as well as its surrounding
homogeneous zone. The so-called homogeneous zone means that, for each of the
regression point, there is a zone in which the studying relationship is spatial stationary.
This zone is abstract and only exists at the mind at this stage.

For every regression point, the GWR examines all data points and calculates their
likelihoods of belonging to its homogeneous zone. According to the First Law of
Geography (Tobler, 1979), this likelihood is determined by the distance between the
regression point and data point. The closer is the distance, the more likely the data point
belongs to the homogeneous zone of the regression point; and vice versa.

Based on these likelihoods, the expected values of the collected data can be obtained at
places of each regression point and its homogeneous zone. Then, the estimations of the
model parameters can be carried out locally at each of the regression point. These steps
are repeated until the estimations are performed for all regression points.

Although there is no special requirement for the location of regression point, this study
performed the GWR at the same positions as the data points.

In statistics, above thoughts are implemented by series of weighted least square (WLS)
(Fotheringham et al., 2002). For a global model,

\[ Y_i = \alpha_0 + \sum_{k=1}^{p} \alpha_k x_{ik} + \theta_i, \quad i = 1, \ldots, n, \]

the OLS estimation is

\[ \hat{\alpha} = (X^T X)^{-1} X^T Y \]

With \( X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \)

\( Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \) and \( \hat{\alpha} = \begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_p \end{bmatrix}. \)
where $\hat{\alpha}$ is the estimation for model parameters, $\alpha^T = (\alpha_0, \alpha_1, \ldots, \alpha_p)$.

For the local model, the model parameters are allowed different at different locations of observation. That is,

$$Y_i = \alpha_{i0} + \sum_{k=1}^{p} a_{ik} x_k + \theta_i, \quad i = 1, 2, \ldots, n,$$

(10)

where the $a_i = (a_{i0}, a_{i1}, a_{i2}, \ldots, a_{ik})^T$ is the vector of model parameter at the location $i$. In this form, the model parameters do not need to be identical at every location of the observation. Therefore the variations in space are allowed.

To estimate the $a_i$, as it stated above, the proximities to the $i$th regression point should be considered for each of the data point. According to these proximities, the weights for every data point could be calculated. This weight shares the same meanings as the above referred “likelihood”, which the closer to the regression point, the higher the weight is added to the data point; and vice versa. With these weights, the local estimation of the model parameters can be then obtained at the location $i$,

$$\hat{\alpha}_i = (X^TW_iX)^{-1}X^TW_iY, \quad i = 1, 2, \ldots, n,$$

(11)

where

$$W_i = \begin{pmatrix} W_{i1} & W_{i2} & \cdots & W_{ij} \\ W_{i2} & \ddots & \cdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ W_{ij} & \cdots & \cdots & W_{in} \end{pmatrix} = \text{diag}(W_{i1}, W_{i2}, \ldots, W_{ij}, \ldots, W_{in})$$

(12)

and $W_{ij}$ denotes the weight added to data point $j$ at the location $i$. 
2.4.2 Weights matrices

Determining weights matrices is necessary for carrying out the local estimations. Fotheringham et al. (2002) suggest several weighting functions, as the Gaussian or near-Gaussian function,

\[ W_{ij} = e^{-\frac{1}{2} \frac{d_{ij}^2}{b^2}} \]  

(13)

the bi-square function,

\[ W_{ij} = \begin{cases} \left[1 - \left(\frac{d_{ij}}{b}\right)^2\right]^2 & \text{if } d_{ij} < b \text{ or } \\
0 & \text{otherwise,} \end{cases} \]

(14)

and the transformation of the bi-square function,

\[ W_{ij} = \begin{cases} \left[1 - \left(\frac{d_{ij}}{b}\right)^2\right]^2 & \text{if } d_{ij} < \tilde{b} \text{ or } \\
0 & \text{otherwise,} \end{cases} \]

(15)

where the \( \tilde{b} \) is the distance between the regression point to the Nth nearest data point.

In these functions, the \( d_{ij} \) denotes the distance between the \( i \)th regression point and the \( j \)th data point, and the \( b \) denotes the bandwidth. The bandwidth is a measure for the distance-decay of the weighting function. The smaller is the \( b \), the larger the extent of the weight decreases along with the distance increasing, and vice versa. In this sense, the value of \( b \) also measures the “the extent to which the resulting local calibration results are smoothed” (Fotheringham et al., 2002).

In addition, the first two functions are examples of the fixed spatial kernel, whereas the third is an instance of the adaptive spatial kernel. Using the adaptive spatial kernel, values of bandwidth are different for different regression points. It varies according to the density of data points in the proximity of the regression point. The higher is the data
points density, the smaller is the value of the bandwidth, and vice versa. It therefore reduces the influences of the distribution of data points on the standard error of the parameter estimations (Fotheringham et al., 2002).

### 2.4.3 Model calibration

Fotheringham et al. (2002) indicate that in practice, the choice of the weighting function doesn’t affect the local estimation. However, for a given weighting function, the value of bandwidth largely influences the result of GWR. So determining the optimal value of bandwidth is an essential part of GWR (for the adaptive spatial kernel, it is the optimal number of data points within the nearest neighborhood of the regression point).

Minimizing the value of the cross-validation (CV) is one of the methods to obtain the optimal bandwidth (Cleveland, 1979; Bowman, 1984; Fotheringham et al., 1998, 2002; Leung, 2000a,b; Huang and Leung, 2002). A score of the form

$$CV = \sum_i \left[y_i - \hat{Y}_{i\omega}(b)\right]^2$$  

(16)

is used. In this expression, $\hat{Y}_{i\omega}(b)$ denotes the fitted value of $Y$ at the location $i$ that obtains using $b$ as the bandwidth but without the data at the regression point. A graph connecting different values of $b$ and $CV$ can be plotted, from which the optimal value $b$ can be found.

Besides, there are other methods for finding the optimal bandwidth, as minimizing the generalized cross-validation criterion (Loader, 1999; Wahda, 1990), the Akaike information criterion (Hurvich et al., 1998), or the Bayesian information criterion (Nakaya, 2002). Although each of these approaches has its merits, considering the aim of this study and the intensity of the relevant computation, the CV was used in this study. So details of these approaches are not reviewed here.

### 2.4.4 Significant tests for spatial variations

As it states by Fotheringham et al. (1998, 2002), Leung et al. (2000a,b), and Huang and Leung (2002), there are two questions should be concerned when applying the GWR. The
The first question is whether the estimation of GWR is significantly better than the OLS estimation for the relationship of concern. The second question is, for each of the variable, whether its parameter significantly varies in space.

Actually, the first question is the goodness-of-fit test for the GWR (Huang and Leung, 2002). The GWR usually provides estimation that more fit to the data than the OLS. However, if this fitness is not significant, the OLS estimation is more preferable and should be used in practice. It is because that the simpler is the model, the easier it can be interpreted and applied. On the other hand, if the GWR estimation is significantly better than the OLS estimation, it is worth to find out which estimated parameters exhibit significant spatial variation over the studying area. It provides a deeper understanding for the studying relationship, and makes it possible to shape more locally pertinent measures and policies.

For these two questions, Fotheringham et al. (1998, 2002) suggest the Monte-Carlo approach. However it is computational intensive and needs more efforts during the programming works. Leung et al. (2000a,b) and Huang and Leung (2002), on the other hand, provide more convenient solutions.

For the first question, the null hypothesis, $H_0$, assumes no significant difference between the GWR and OLS estimations. Then the statistic $F_1$ is calculated.

$$F_1 = \frac{\frac{RSS_g}{\delta_i}}{\frac{RSS_o}{(n-P-1)}},$$

where

$$RSS_g = Y^T(I-L)(I-L)^TY,$$

$$RSS_o = Y^T(I-X(X^TX)^{-1}X^T)Y,$$

$$\delta_i = Tr [(I-L)^T(I-L)^T],$$

where $i=1,2$. 

(17)
\[
L = \left\{ \begin{array}{c}
X_1^T \left[ X^T W_1 X \right]^{-1} X^T W_1 \\
X_2^T \left[ X^T W_2 X \right]^{-1} X^T W_2 \\
\vdots \\
X_n^T \left[ X^T W_n X \right]^{-1} X^T W_n
\end{array} \right. ,
\]
(21)

\[
X_i^T = \left( 1, x_{i1}, \ldots, x_{ip} \right)
\]
(22)

The distribution of \( F_1 \) is an \( F \)-distribution with degrees of freedom \( (\delta_1^2 / \delta_2, n - P - 1) \). If \( F_1 < F_{1, \alpha} (\delta_1^2 / \delta_2, n - P - 1) \), then the Ho is rejected at the significant level \( \alpha \), and the GWR estimation is concluded better than the OLS estimation in describing the relationship of concern. Otherwise, at the significant level \( \alpha \), the GWR and OLS estimation have no difference.

For the second question, the null hypothesis, \( H_0 \) states that there is no spatial variation, is \( \hat{\alpha}_{1k} = \hat{\alpha}_{2k} = \ldots = \hat{\alpha}_{nk} \), \( (k = 0, 1, 2, \ldots, P) \), and the alternative hypothesis, \( H_1 \), is not all \( a_{nk}^* \) are equal. Then construct the statistic \( F_3(k) \),

\[
F_3(k) = \frac{V_k^2 / \gamma_1}{\sigma^2},
\]
(23)

where

\[
V_k^2 = \frac{1}{n} \hat{\alpha}(k)^T \left( I - \frac{1}{n} \right) \hat{\alpha}(k),
\]
(24)

\[
\gamma_i = \text{tr} \left[ \frac{1}{n} B^T \left( I - \frac{1}{n} \right) B \right], \quad i = 1, 2,
\]
(25)
\[
B = \begin{bmatrix}
    e_k^T \left[ X^T W_1 X \right]^{-1} X^T W_1 \\
    e_k^T \left[ X^T W_2 X \right]^{-1} X^T W_2 \\
    \vdots \\
    e_k^T \left[ X^T W_n X \right]^{-1} X^T W_n 
\end{bmatrix}
\]

(26)

\[
\hat{\sigma}^2 = \frac{RSS}{\delta_t}.
\]

(27)

In these expressions, the \( I \) is the identity matrix with order of \( n \), \( J \) is an \( n \times n \) matrix that each element equals to one, and \( e_k^T \) is row vector that all elements are zero except the \((k+1)\)th element that equals to one.

Approximately, the distribution of \( F_3(k) \) is the \( F \)-distribution with degrees of freedom \( (\gamma_1^2/\gamma_2, \delta_1^2/\delta_2) \). If \( F_3(k) > F_{\alpha}(\gamma_1^2/\gamma_2, \delta_1^2/\delta_2) \), then the \( H_0 \) is rejected at the significant level \( \alpha \), and \( a_k^* \) is concluded spatially varies across the studying area. Otherwise, \( H_1 \) is accepted and \( a_k^* \) is considered stationary in space.

### 2.5 Some statistical tests

Assume we have a linear regression model,

\[
Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_k X_k + \theta.
\]

(32)

Then, the following tests can be carried out.

#### 2.5.1 t-test

To test the null hypotheses \( H_0 \) that the actual value a parameter in a regression model is zero, the t-test was used in this study. That is to test if

\[
a_k = 0.
\]

(33)
The variable \( t \),

\[
    t = \frac{\hat{a}_k}{SE(\hat{a}_k)}.
\]  

(34)

is \( t \)-distributed with the degree of freedom \((n-P-1)\). If \( t > t_\alpha(n-P-1) \), then \( H_0 \) is rejected at the significant level \( \alpha \), and the value of \( a_k \) is concluded different to zero. Otherwise, at the significant level \( \alpha \), the \( a_k \) equals to zero.

2.5.2 \( F \)-test

Usually, some groups of variables in a regression model are suspected having no obvious influence on the dependent variables. That is, say for the equation (32), suspecting

\[
    a_{k-2} = 0; \ a_{k-1} = 0; \ a_k = 0.
\]  

(35)

The \( F \)-test, which actually tests if two estimated variances are equal, could be used in this circumstance. The \( F \) statistic is

\[
    F = \frac{(RSS_r - RSS_{wr})/q}{RSS_{wr}/(n - P - 1)},
\]  

(36)

where \( RSS_r \) and \( RSS_{wr} \) are the sums of squared residual of the model with and without \( X_{k-2}, X_{k-1}, X_k \), respectively, and the \( q \) is number of restricted parameters. The null hypothesis, \( H_0 \), of \( F \)-test is that these two estimated variances are equal. If that is the case, these two regression models actually have no significant difference, and variables \( X_{k-2}, X_{k-1}, X_k \), could therefore be excluded from the unrestricted model.

The \( F \) variable is \( F \)-distributed with degrees of freedom \((q, n-P-1)\). If \( F > F_\alpha(q, n-P-1) \), then the \( H_0 \) is rejected at the significant level \( \alpha \). It can be concluded that the restricted and unrestricted models have a significant difference. So not all restricted variables can be excluded from the unrestricted model. Otherwise, at the significant level \( \alpha \), these two
models have no significant difference and all these restricted variables \( X_{k-2}, X_{k-1}, X_{k} \),
could therefore be excluded from the unrestricted model.

2.5.3 White Test

The White test is a test for heteroskedasticity. Heteroskedasticity means that, the variance
of the unobservable factor, \( \theta \), is not constant if the value of independent variable changes.
The problem of the heteroskedasticity does not make the estimation biased or inconsistent.
However it makes the \( t \)-value is no longer \( t \)-distributed and \( F \)-value is no longer \( F \)-
distributed, and therefore the results of these statistical tests become unreliable.

To perform the White test, following regression model is required.

\[
\hat{\mu}^2 = \delta_0 + \delta_1 \hat{Y} + \delta_2 \hat{Y}^2 + \text{error},
\]

(37)

where the \( \hat{\mu} \) is residual and the \( \hat{Y} \) is the fitted value from OLS estimation of the equation
(32). \( F \)-test is then performed to test the null hypotheses \( H_0 \) that \( \delta_1 = 0; \delta_2 = 0 \). If the
value of the \( F \) statistic is not larger than \( F_{\alpha}(2, n-3) \), then it could be concluded that no
heteroskedasticity exists at the significant level \( \alpha \). Otherwise, at the significant level \( \alpha \),
the problem of the heteroskedasticity can not be ignored and special measures should be
carried out.

2.5.4 RESET test

The RESET test examines the problem of model misspecification. Model
misspecification occurs when a variable should be modeled in its quadratic or logarithmic
form instead of its linear form. The problem of model misspecification may cause effects
similar to the problem of the omitted variables, since it actually excludes one or more
relevant variables from the regression model.

To perform the RESET test, the following regression model is established.

\[
Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_K X_K + \delta_1 \hat{Y}^2 + \delta_2 \hat{Y}^3 + \text{error},
\]

(38)
where the $\hat{Y}$ is the fitted value from OLS estimation of the equation (32). $F$-test is then performed to test the null hypotheses $H_0$ that $\delta_1 = 0; \delta_2 = 0$. If the value of the $F$ statistic is not larger than $F_{\alpha}(2, n-k-3)$, then it could be concluded at the significant level $\alpha$ that the model is properly specified. Otherwise, at the significant level $\alpha$, the model is suffering the misspecification problem.
3 Materials and methods

3.1 Materials and data processing

For this study, the Uppsala diocese provided several datasets from the church statistics in 2002-2006 and a church map in Shapefile format. These datasets included demographic data as well as the statistics for membership and church services for each parish, and economic data of church operations for each pastorat.

In the Church of Sweden, a pastorat usually consists of many parishes, and economic data of church operations are only available at the pastorat level. Other data provided at the parish level were thereby aggregated into the pastorat level.

However, the delineation of these pastorats was not all the same in this period. Names of pastorats had some slight changes as well. For example, the number of pastorats in 2006 was 70; however it had been 75 in 2004. Besides, the parish code was different between years before 2005 and after 2005. To cope with these differences, this study used the pastorats in 2006 as the standard unit. Data from years previous to 2006 were then adjusted accordingly. In this process, it was found that data of 2002 was quite different to other, and it was difficult to make adjustment without help of more information. So this study didn’t include data of this year. Microsoft EXCEL 2003, VBA and ESRI ArcGIS 9 were employed during these works.

This study concerned the church performance. Although the link between the Church of Sweden and the State was removed, and newborn didn’t automatically have the status of membership anymore, the rate of membership was still very high in each pastorat, and wasn’t a proper index for the church performance (Hamberg and Pettersson, 1994). Therefore, another measurement, the Weekly Attendance Score (WAS) (Hamberg and Pettersson, 1994; Berggren, 1997) was employed in this study. The WAS, as its definition, is the average percentage of population that attending services per week.

The key independent variables of this study were the total numbers of kinds of church services, and the relevant operation cost. The classification of the church services was based on the dataset NyckeltalstabellerUppsala_2006. Hence church services were grouped into main church services (huvudgudstjänster), church actions (kyrkliga handlingar), and other church services (andra gudstjänster).
For the cost of church operation, several kinds of cost were considered. These are the total operation cost ($CST$), the cost for pastorat and funeral services ($PAS$), the cost for the real estate management ($RE$), and the cost for the general administration ($GNLAD$). For making them compatible to the $WAS$, the costs and services were adjusted to values per capita. Besides, the ratio of population with age above 65, under 19, and the membership rate were included also as the control variables (Hamberg and Pettersson, 1994; Berggren, 1997).

### 3.2 Models

In order to model the variation in the $WAS$, the following regression model was applied:

$$WAS_i = a_0 + a_1R19_i + a_2R65_i + a_3MR + a_4ROTH_i + a_5RMA_i + a_6RACT_i + a_7COST_i + a_8COST_i \ast ROTH_i + a_9 COST_i \ast RMA_i + a_{10} COST_i \ast RACT_i + \theta_i,$$  \hspace{1cm} (39)

where,

- $WAS$ was the weekly attendance score;
- $R19$ and $R65$ were the ratio of population with age under 19 and above 65, respectively;
- $MR$ was the membership rate;
- $ROTH$, $RMA$, and $RACT$ were the numbers of other church services, main church services, and church actions supplied for every 100 persons, respectively;
- $COST$ was the church operation cost per capita. (Since the discussion about which cost variables should be used hadn’t been performed, the symbol $COST$ was temporarily used. In this study, the $COST$ could be the total operation cost per capita ($CST$) and the cost for pastorat activities per capita ($PAS$). Analysis about the cost for the real estate management per capita ($RE$), and the cost for the general administration per capita ($GNLAD$) was performed later as well)

The estimations of total operation cost per capita ($CST$) and the cost for pastorat activities ($PAS$) were performed later.

Using the similar symbols, the Y-lag and FD models were therefore,

$$WAS_i = a_0 + a_1R19_i + a_2R65_i + a_3MR + a_4ROTH_i + a_5RMA_i + a_6RACT_i + a_7COST_i + \ldots$$
\[ a_8 \text{COST}_i \times \text{ROTH}_i + a_9 \text{COST}_i \times \text{RMA}_i + a_{10} \text{COST}_i \times \text{RACT}_i + a_{11} \text{WAS\_LAG}_i + \theta_i, \]

and

\[ \Delta \text{WAS}_i = a_0 + a_1 \Delta \text{R19}_i + a_2 \Delta \text{R65}_i + a_3 \Delta \text{MR} + a_4 \Delta \text{ROTH}_i + a_5 \Delta \text{RMA}_i + a_6 \Delta \text{RACT}_i + a_7 \Delta \text{COST}_i + a_8 \Delta (\text{COST}_i \times \text{ROTH}_i) + a_9 \Delta (\text{COST}_i \times \text{RMA}_i) + a_{10} \Delta (\text{COST}_i \times \text{RACT}_i) + \theta_i, \]

where \( \text{WAS\_LAG} \) denoted the lagged \( \text{WAS} \), and the “\( \Delta \)” denoted the difference between two years.

### 3.3 Modeling strategies comparison and GWR estimation

A main task of this study was to choose a modeling strategy from the simple regression model (Eq. 39), \( Y\)-lag model (Eq. 40), and FD model (Eq. 41). The choice was made by comparing the OLS estimates for these models using the data of 2004 and 2006 in the sense of the magnitude, sign, and statistical significance.

After the modeling method had been selected, sets of statistical tests were made to guide the model calibration. It included the heteroskedasticity test, model misspecification test, redundant variables test, and tests to choose variables representing the \( \text{COST} \) and lagged-\( \text{WAS} \). In this study, the White test suggested by Woodbridge (2002) that uses the fitted-value of \( Y \) and the RESET test (Ramsey, 1969) were employed for testing the heteroskedasticity and the model misspecification. For the problems of redundant variables and variables choose, the \( F \)-test and the adjusted \( R^2 \) were used. The econometric software Eviews 5.0 was employed for these analyses.

When carrying out the GWR estimation, the near-Gaussian equation (13) was used as the weighting function, and the approach of minimizing the cross-validation score (16) was employed to obtain the optimal bandwidth. Leung’s \( F_1 / F_3 \) tests (Leung et al. 2000a,b; Huang and Leung, 2002) were also performed to test the spatial variation. Microsoft EXCEL 2003 and VBA were applied to carry out these calculations. Relevant codes were reported in the Appendix. ESRI ArcGIS 9.1 was used as the GIS platform, and Anselin’s
GeoDa 0.9.5 (2005) were also applied to calculate the spatial data as coordinates of polygon centroid and the value of Moran’s $I$. 
4 Results and analysis

This section presented results of parameter estimation for the regression models mentioned in chapter 3, which include the simple regression model (Eq. 39), Y-lag model (Eq. 40), and First-differencing model (Eq. 41). The results were then analyzed with respect to the three modeling strategies. The result of the GWR estimation was also analyzed along with Leung’s $F_1 / F_3$ tests and a comparison with the results obtained from these OLS regressions. It helped to compare the GWR and OLS.

4.1 Performance of the non-geographical regression models

The objective of this subsection is to compare the regression models without considering geographic factors. Table 1 describes the OLS estimations for these models using data of 2004 and 2006.
Table 1: Estimation of the parameters of the simple regression, Y-lag, and FD models for data of 2004 and 2006. Numbers in brackets are *p-values* for the corresponding parameter estimations.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Original</th>
<th>Y-lag</th>
<th>FD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R19_{-6}$</td>
<td>0.046922</td>
<td>-0.011025</td>
<td>0.028015</td>
</tr>
<tr>
<td></td>
<td>[0.8516]</td>
<td>[0.8131]</td>
<td>[0.6937]</td>
</tr>
<tr>
<td>$R35_{+6}$</td>
<td>0.047373</td>
<td>0.003996</td>
<td>0.101698</td>
</tr>
<tr>
<td></td>
<td>[0.4262]</td>
<td>[0.9263]</td>
<td>[0.3665]</td>
</tr>
<tr>
<td>$MR6$</td>
<td>-0.001691</td>
<td>-0.013079</td>
<td>0.07048</td>
</tr>
<tr>
<td></td>
<td>[0.9514]</td>
<td>[0.5148]</td>
<td>[0.4767]</td>
</tr>
<tr>
<td>$ROTH6$</td>
<td>-0.527462</td>
<td>-0.530818</td>
<td>0.35297</td>
</tr>
<tr>
<td></td>
<td>[0.0128]</td>
<td>[0.0027]</td>
<td>[0.6921]</td>
</tr>
<tr>
<td>$RMA6$</td>
<td>3.847166</td>
<td>2.665601</td>
<td>0.933252</td>
</tr>
<tr>
<td></td>
<td>[0.0003]</td>
<td>[0.0007]</td>
<td>[0.1014]</td>
</tr>
<tr>
<td>$RACT6$</td>
<td>-0.529394</td>
<td>-0.377777</td>
<td>1.745782</td>
</tr>
<tr>
<td></td>
<td>[0.6733]</td>
<td>[0.6764]</td>
<td>[0.0371]</td>
</tr>
<tr>
<td>$CST6$</td>
<td>0.001134</td>
<td>0.000827</td>
<td>0.001686</td>
</tr>
<tr>
<td></td>
<td>[0.3122]</td>
<td>[0.3069]</td>
<td>[0.0484]</td>
</tr>
<tr>
<td>$CST6*ROTH6$</td>
<td>0.003557</td>
<td>0.000322</td>
<td>0.000386</td>
</tr>
<tr>
<td></td>
<td>[0.0014]</td>
<td>[0.0011]</td>
<td>[0.6392]</td>
</tr>
<tr>
<td>$CST6*RMA6$</td>
<td>-0.001459</td>
<td>-0.0011</td>
<td>-0.00209</td>
</tr>
<tr>
<td></td>
<td>[0.0027]</td>
<td>[0.002]</td>
<td>[0.5938]</td>
</tr>
<tr>
<td>$CST6*RACT6$</td>
<td>0.000468</td>
<td>0.000252</td>
<td>-0.000501</td>
</tr>
<tr>
<td></td>
<td>[0.4245]</td>
<td>[0.3516]</td>
<td>[0.1995]</td>
</tr>
<tr>
<td>$C$</td>
<td>-1.190199</td>
<td>0.156155</td>
<td>-0.21732</td>
</tr>
<tr>
<td></td>
<td>[0.7291]</td>
<td>[0.9498]</td>
<td>[0.3253]</td>
</tr>
<tr>
<td>$WAS$</td>
<td>[0]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td>70</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>$P$</td>
<td>11</td>
<td>12</td>
<td>11</td>
</tr>
</tbody>
</table>

The largest probability that the null hypothesis of a statistical test cannot be rejected is measured by the *p-value* (Woodbridge, 2002). It is usually computed automatically by the statistical software when carrying out the regression analysis. If the *p-value* is smaller than the pre-set level of the statistical significance, then the estimation of this parameter is considered as statistically significant; and vice versa.

This study set the significant level to 15%. As a result, the estimations those have their *p-value* smaller than 0.15 were considered as statistically significant during this study. This is not a strong significant level but still made sense, since the degrees of freedom of models developed in this study were slightly less than 60. However, if the sample size is increased, the significant level might also be decreased, say, to 5%.

Columns “Original” and “Y-lag” in Table 1 show results of the OLS estimation for models of original and Y-lag, separately. By comparing them, two important results can be noticed:
First, when including the Y-lag, the model’s ability of explanation was improved. The adjusted $R^2$ for the Y-lag was about 85%, which was larger than the one for the simple regression model, 71%. It meant that more variations in the weekly attendance score across parishes in Uppsala diocese at the year of 2006 was caught by the Y-lag model than the simple regression model.

The result of the Y-lag model was not a surprise. The lag of Y in time usually has some inertia effects on the present Y. In the Y-lag model of this study, the estimated parameter of WAS4 reached 0.59, which implied if two parishes differed in one unit of weekly attendance score (WAS) in 2004, and if the two parishes took the same church operations in 2006, then they would have the WAS different in 0.59 unit at this year. In addition, the $t$-value for this parameter estimation was high up to 7.46, which corresponded to an almost zero $p$-value. It denoted that their statistical correlation was very significant.

Second, statistically significant estimations for parameters were similar in both the simple regression model and the Y-lag model. Estimations for parameter of key variables, which were the rates of church services supplies per 100 persons ($ROTH$, $RMA$, $RACT$), the church operation cost per capita ($CST$), and their interactions, were similar in signs, magnitudes, and statistical significances. Among these statistically significant estimations, parameter for the rate of main church services per 100 persons ($RMA$) was mostly differently estimated in these models. After including the Y-lag, it reduced 35% in the magnitude. However, both of the estimations were in the 85% confidence intervals of each other. Estimations for demographic variables, the ratios of population with age below 19 and above 65 ($R19$, $R65$), and the membership rate ($MR$), had larger differences in these two models. However, all of them were statistically insignificant in both models.

By jointly considering the details of the regression analysis, it was found that the $WAS4$ was significantly correlated to other dependent variable of the model. However, including $WAS4$ to the model could not bring any significant changes to the estimation of the model parameters. According to the Gauss-Markov assumptions behind OLS (Woodbridge, 2002), this situation could happen in two cases. The first case was that the simple regression model was well defined, so the parameter estimates of these already included variables were unbiased. Adding a relevant variable could not significantly influence the estimations for these variables but only be able to improve a model’s overall explanation ability or to increase its value of the $R^2$ or the adjusted $R^2$. 


The second explanation is that the simple regression model suffered the problem of omitted variables. By including the Y-lag, this problem cannot be solved. It could happen when the Y-lag is not correlated to the independent variables. Adding it into the regression model could only increase the $R^2$ or the adjusted $R^2$, but has no effect on the problem of omitted variables.

Obviously the first situation was more preferable. However, as it was going to be shown below, the latter case seems to be at hand.

In both of the original and Y-lag models, some parameter estimations were not as expected. Taking the simple regression model for an instance, the parameter of variables $ROTH$ and $ROTH^CST$ measured the partial/marginal effect of the supply of “other church services” on WAS, considering the value of the church operation cost per capita. The partial/marginal effect of a variable is the contribution of one unit increasing of this variable to the dependant variable when hold all other independent variables fixed (Woodbridge, 2002). They were one positive and one negative, and both statistically significant. Similarly, the parameters of $RMA$ and $RMA^CST$ expressed the marginal effect of the supply of church main services at a given level of the church operation cost per capita. They were also statistically significantly one positive and one negative. However, their ways of these one positive and one negative pattern were reversed. It implied that the supplies of the church other services ($ROTH$) and church main services ($RMA$) affected the WAS differently. Along with the increasing the level of church operation cost per capita ($CST$), the marginal effect of $ROTH$ increased, while the marginal effect of $RMA$ decreased at the same time. It was a little bit counterintuitive; since the common idea for this issue was that they may differ in magnitude, but should be in a same pattern

This counterintuitive result may suggest the problem of omitted variable (Woodbridge, 2002). But it could also be an improvement for the understanding of the pastoral operations. So far, estimations for the original and Y-lag models were compared. But more information was required to make it clear whether these models suffered the problem of the omitted variables, which may cause the parameter estimations to be biased.

More information could be obtained by comparing the original and Y-lag models with the FD model. As it state above in the section 2.1, although the FD model in fact measures the relationship between changes of variables, its parameters can also be explained as these variables are in the original or Y-lag models. It therefore made it possible to
compare the parameters in these models. However, the $R^2$ or the adjusted $R^2$ of the FD model can not be compared to those of the original and Y-lag models. It is because that the $R^2$ or the adjusted $R^2$ in the FD model explains how much variation of the $\Delta Y$ is caught by the independent variables, whereas in the simple regression model they describe how much variations of the Y is explained by the explanatory variables.

In the FD model, estimations for parameters of key variables, $ROTH$, $RMA$, $RACT$, and $CST$, were all statistically significant at the level of 15%. However, these estimations were essentially different to those of the original and Y-lag model. They differed both in magnitudes and in signs. For the parameters of the FD model the same explanations as those of the simple regression model can be applied. So the large differences between these statistically significant estimations suggest the problem of the omitted variables for one or both of the models (Woodbridge, 2002). Considering that the extent of the omitted variables problem had already been reduced here in FD model, whereas the no measures had been applied to the simple regression model, it can be inferred that the simple regression model did suffer the omitted variables problem. If it was the case for the simple regression model, then the Y-lag model that had the similar results as the simple regression model had this type of problem also. These models therefore produced the biased estimations probably for they missed to control the unobserved factors that had correlations with the independent variables.

By a set of comparisons two out of three potential choices for modeling were excluded and left the first-differencing estimation as the most suitable one. It was therefore employed by this study as the modeling strategy.

### 4.2 Selection of variables

After choosing the first-differencing as the modeling strategy, several tests could then be performed to detect potential problems of the model specification. These problems could be the heteroskedasticity, model misspecification, and inclusion of redundant variables.

Besides, the total cost per capita ($CST$) was so far used to represent the cost of the operations. However, other cost measures such as the cost of pastorat activities per capita ($PAS$) were also available. A comparison of these cost measures was therefore required.
Table 2: Statistical tests for the choice of the COST variable and the problem of redundant variables. Numbers in bracket are *p*-values for the corresponding tests.

<table>
<thead>
<tr>
<th>Tests</th>
<th>CST</th>
<th>PAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of significant variables</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.61676</td>
<td>0.627685</td>
</tr>
<tr>
<td>White test (<em>p</em>-value)</td>
<td>0.117022</td>
<td>0.296749</td>
</tr>
<tr>
<td>RESET test (<em>p</em>-value)</td>
<td>0.873816</td>
<td>0.884304</td>
</tr>
<tr>
<td>$F$-test (Demo)</td>
<td>0.77421</td>
<td>[0.51144]</td>
</tr>
<tr>
<td>$F$-test (RE=GNIAD)</td>
<td>0.528956</td>
<td>[0.000009]</td>
</tr>
<tr>
<td>$F$-test (ROTH)</td>
<td>36.38323</td>
<td>[0.01437]</td>
</tr>
<tr>
<td>$F$-test (RIH)</td>
<td>4.574684</td>
<td>[0.02156]</td>
</tr>
</tbody>
</table>

The first four rows in Table 2 shows the results of the comparison between the use of the total cost per capita (*CST*) and the cost of pastorat activities per capita (*PAS*) as an independent variable. First, there were five significant variables in the model using *PAS*, whereas three for the model using *CST*. Second, although the difference was not so large, using the *PAS* improved the model explanation ability. The value of adjusted $R^2$ increased from 0.617 in the model using *CST* to 0.628 in the model using *PAS*. Third, the results of the White test showed that the model with *PAS* was free of the heteroskedasticity problem at the significant level of 15%. Its *p*-value, which denotes the possibility that a model is not suffering the problem of heteroskedasticity, was 28.7%. However, the *p*-value of the White test for the model with *CST* was 11.7%, which indicates the existence of the heteroskedasticity problem of this model. In addition, the hypothesis of the model misspecification was significantly rejected for both of these models by the Ramsey’s *RESET* test (1969). The *p*-values that denote the possibility that a model is properly specified were both high up to 80% for these two models.

These facts indicated that the cost of pastorat activities per capita (*PAS*) was a better indicator for the church operation cost than the total cost per capita (*CST*) for this study. The *PAS* was therefore selected as the cost indicator during this study.

The rest part of the Table 2 shows results of a serial of F-tests. There F-tests were performed to detect the problem of redundant variables within the FD model with *PAS*. At first, combinations of church services supplies and their interaction with the church operation cost were jointly significant in statistics. It implied that it did make sense to include them to the model. On the other hand, demographic variables (*Demo* in Table 2)
that included the ratios of population with age below 19 and above 65 \((R19, R65\) in Table 2), and the membership rate \((MR)\), were jointly statistically insignificant. Therefore, all these three variables could be removed from the FD model at the same time. It was not surprising since within a two-year time span, the demographic structure in parishes might change, but only in a limited extent. Besides, the variables of the cost in real estate management per capita \((RE)\) and the cost of general administration per capita \((GNLAD)\) were jointly statistically insignificant. It suggested that the investments to the real estate management and general administration had no significant influences on the WAS at that year. Therefore, demographic variables \((Demo)\), the cost in real estate management per capita \((RE)\), and the cost of general administration per capita \((GNLAD)\) could be excluded from the FD model.

All above discussion resulted in the model shown in the Table 3.

Table 3: The estimation for the calibrated FD model

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient</th>
<th>t-Value</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROTH1-ROTH4</td>
<td>0.766502</td>
<td>3.894504</td>
<td>0.0002</td>
</tr>
<tr>
<td>RIA6-EIA4</td>
<td>-0.394694</td>
<td>-0.615893</td>
<td>0.5402</td>
</tr>
<tr>
<td>RACT6-RACT4</td>
<td>2.007958</td>
<td>2.956556</td>
<td>0.0044</td>
</tr>
<tr>
<td>PAS6-PAS4</td>
<td>0.001661</td>
<td>1.023195</td>
<td>0.3102</td>
</tr>
<tr>
<td>PAS6+ROTH6-PAS4+ROTH4</td>
<td>-0.000252</td>
<td>-1.847178</td>
<td>0.0695</td>
</tr>
<tr>
<td>PAS6+RIA6-PAS4+RIA4</td>
<td>0.000801</td>
<td>1.620751</td>
<td>0.1101</td>
</tr>
<tr>
<td>PAS6+RACT6-PAS4+RACT4</td>
<td>-0.001136</td>
<td>-2.087781</td>
<td>0.0485</td>
</tr>
<tr>
<td>C</td>
<td>-0.228008</td>
<td>-3.266506</td>
<td>0.0015</td>
</tr>
<tr>
<td>Adjusted (R^2)</td>
<td>0.637644</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White test ((p-value))</td>
<td>0.213197</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RESET test ((p-value))</td>
<td>0.611331</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.3 **GWR estimations**

The results of the GWR estimations are presented in Figures 2 to 4. Figure 1 shows spatial distributions of the residual from the estimations for the global model.
Figure 1: Spatial distributions of residuals from estimations for global models. The upper left one was for the simple regression model, the upper right one was for the Y-lag model, and the lower one was for the FD model.

Determining the optimal value of bandwidth was crucial for the GWR estimations. Figure 2 showed optimal values of bandwidth for the original, Y-lag, and FD models using the approach of minimizing the CV score.
Figure 2: Optimal values of bandwidth for the original, Y-lag, and FD models. The X-axis denoted the value of the bandwidth, and Y-axis denoted the CV score.

For comparing the OLS and GWR estimations, the most significant estimations of the partial effects of ROTH as well as their difference were presented in Figures 3 and 4. The results from Leung’s $F_1/F_3$ tests are presented in Table 4.

Figure 3: Marginal effects of ROTH estimated by OLS and GWR.
Figure 4: Differences between Marginal effects of ROTH estimated by OLS and GWR.
Table 4: Results of the Leung’s $F_1$ and $F_3$ tests for spatial variations. Numbers in “[]” were p-values for the corresponding parameter estimations, which denoted the possibility that a parameter is constant across the space.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Origin</th>
<th>Y-Lag</th>
<th>FD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>0.791759402</td>
<td>0.917106738</td>
<td>0.976257418</td>
</tr>
<tr>
<td>$F_3(C)$</td>
<td>[0.799355]</td>
<td>[0.624611]</td>
<td>[0.535592]</td>
</tr>
<tr>
<td>$F_3(R19)$</td>
<td>1.212393566</td>
<td>0.3869734</td>
<td>0.297711534</td>
</tr>
<tr>
<td>$F_3(R65)$</td>
<td>0.649406897</td>
<td>1.461545855</td>
<td>4.788738556</td>
</tr>
<tr>
<td>$F_3(MR)$</td>
<td>6.872226838</td>
<td>4.683163948</td>
<td>0.211098265</td>
</tr>
<tr>
<td>$F_3(ROTH)$</td>
<td>[0.011569]</td>
<td>[0.034902]</td>
<td>[0.080161]</td>
</tr>
<tr>
<td>$F_3(MA)$</td>
<td>1.393497037</td>
<td>1.273396356</td>
<td>1.353238438</td>
</tr>
<tr>
<td>$F_3(ACI)$</td>
<td>1.586962257</td>
<td>1.134936311</td>
<td>1.808610454</td>
</tr>
<tr>
<td>$F_3(RE)$</td>
<td>0.916947654</td>
<td>5.595205883</td>
<td>0.111308442</td>
</tr>
<tr>
<td>$F_3(GNLAD)$</td>
<td>1.739695023</td>
<td>0.288034058</td>
<td>0.817413998</td>
</tr>
<tr>
<td>$F_3(PAS)$</td>
<td>0.130966025</td>
<td>0.485040841</td>
<td>0.895543034</td>
</tr>
<tr>
<td>$F_3(PAS*ROTH)$</td>
<td>0.602433773</td>
<td>0.433090017</td>
<td>1.93875517</td>
</tr>
<tr>
<td>$F_3(PAS*RMA)$</td>
<td>1.151013637</td>
<td>1.217412798</td>
<td>0.181119791</td>
</tr>
<tr>
<td>$F_3(PAS*RACT)$</td>
<td>1.468864485</td>
<td>0.767455209</td>
<td>2.127908995</td>
</tr>
<tr>
<td>$F_3(WAS4)$</td>
<td>0.334285587</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5 shows the spatial patterns of the church service supply. Here, the rates of church services supplies per 100 persons ($ROTH$, $RMA$, and $RACT$), as well as the cost of pastorat activities per capita ($PAS$) at 2006 are shown.
4.4 Comparison between OLS and GWR

Cliff and Ord (1973, 1981) states that analyzing the degree of spatial autocorrelation in the residual could help when considering if a global model is suitable for a spatial data set. If there are significant spatial autocorrelations in the residual, that is the magnitude and sign of the residual is related to the location of observation, the statistical inferences of
the global model would suffer problems. Then there are grounds for applying the GWR, which could largely release these problems.

Figure 1 shows the distributions of residuals of model estimations in the study area. Visually, no obvious spatial pattern was spotted, which suggested no spatial autocorrelations existing among them. A statistic test, the Moran’s I test (Anselin, 1995, 1998) which measures the degree of spatial autocorrelation, supported this judgment. As it defines, the value of Moran’s I varies between minus one and positive one. If no spatial autocorrelation presents, its value equals to \(-1 / (n-1)\) (Wong and Lee, 2005). In this study, the Moran’s I were 0.0338, -0.0106, and -0.0294 respectively for the original, Y-lag, and FD models. They were quite close to the value of -0.0145 that is the value if no spatial autocorrelation presents in this study. It suggested almost no spatial autocorrelation among the residuals of these models. In this sense, there was no need to employ the complicated GWR.

Another reason for using the GWR is the spatial heteroskedasticity, which means that there may be spatial non-stationarity in the relationship of concern. In this study, the GWR was applied to obtain local estimations for the original, Y-lag, and FD models, as well as the statistics, the \(F_1\) and \(F_3\) (Leung et al., 2000a,b; Huang and Leung, 2002), for testing their significances of spatial variations. The result of the \(F_1\) test indicated (Table 4) that, for the relationship between the church services supply and the pastoral performance in the Uppsala diocese, there were no statistically significant differences between the OLS and GWR estimations. However, on the other hand, the results of the \(F_3\) test pointed out some significant variations existing in the parameter estimations for all three models.

For the simple regression model, parameters for the interception and the membership rate (MR) varied significantly across the studying area. Correspondingly for the Y-lag model, the parameters for MR and the cost in real estate management per capita (RE) were significantly spatial non-stationary. However, while they had statistically significant variations across the space, their estimates were insignificant as such in OLS. Among these estimations, the most significant estimate was the parameter of MR in the Y-lag model. But its p-value was high up to 43.3%. For the basic thought of the \(F_1\) test is comparing the standard errors of GWR and OLS estimations (Leung et al., 2000a,b; Huang and Leung, 2002), and the contribution of the insignificant estimation to the standard error is quite small (Woodbridge, 2002), these significant spatial variations didn’t conflict with the conclusion of the \(F_1\) test.
For the FD model, parameters for the ratio of population with age above 65 ($R65$) and the supply of other church services ($ROTH$) showed significant spatial variations. The estimation for the parameter of $R65$ was insignificant in OLS with a $p$-value of 54.2%. As discussed above, it didn’t damage the conclusion of the $F_1$ test. But, for another variable, the $ROTH$, its parameter had a significant OLS estimation. $ROTH$ was one of the key variables in this study, and the magnitude of its parameter reached the 0.81. This result posed a question: how to explain this conflict between the $F_1$ and $F_3$ tests?

Obviously, the technical discussion about this question would depart from the direction of this study. Instead, an alternative treatment was keeping the results of the OLS and GWR estimations of the FD model at the same time. In this study, the estimation of the partial effect of $ROTH$ was, for example, the $(a_4^*+a_{10^*}PAS)$ instead of the pure $a_4^*$, where $a_4^*$ and $a_{10^*}$ were estimations for parameters of $ROTH$ and $ROTH \cdot PAS$. Therefore, it could make sense that comparing OLS and GWR estimations of actual partial effects of $ROTH$ at the studying area. If the results were similar, then in a practical manner, there were no obvious differences between the OLS and GWR estimations of the FD model. Considering the principle of the simpler the better, the OLS estimation for the FD model should be applied. On the other hand, if the results were largely different, and the difference had some specific spatial patterns, then the GWR estimation for the FD model should be used.

Figure 3 describes the distribution of partial effects of the $ROTH$ in the studying area. These two maps used the same breaking values for an easy comparison. Visually, they were almost the same besides some slight differences. It supported the conclusion of the $F_1$ test that there were no significant differences between the OLS and GWR estimations. This study mapped the differences as well. Figure 4 indicates that the differences exhibited some spatial patterns, which GWR intended to overestimate the marginal effect of $ROTH$ at the northern part of the Uppsala diocese whereas the OLS overestimated it at the southern part of the Uppsala diocese. This result supported the conclusion of the $F_3$ test. However, this support was weak, since the magnitudes of the differences were quite small, and in practice these differences can be ignored. All in all, for this study, the relationships of concern were spatial stationary and the OLS estimation of the FD model was competent.
4.5 **Model interpretation**

The discussions above resulted in the model shown in the Table 3. This model included the interaction terms of the church services supplies and the cost of pastoral activities per capita ($PAS$). So the marginal effects of the church services supplies were influenced by the magnitude of $PAS$. In addition, for the same reason, the judgment about the zero marginal effect can not be made solely based on the result of the $t$-test. For example, in this study, the parameter of the church main service supply per capita ($RMA$) was negative and statistically insignificant. However, on the other hand, $RMA$ and its interaction term with the $PAS$ were jointly statistically significant with a large value of the $F$-test, 6.96. The corresponding $p$-value was 0.002, strongly rejected the hypothesis that both of them had no influences on the $WAS$.

To explain the marginal effects of the church services supplies, the levels of the cost of pastoral activities per capita ($PAS$) should be set at first. One way was to use the average values, that is, 1130.47 for 2004 and 1194.57 for 2006. With these values, in 2006 the partial contributions of church services supplies ($ROTH$, $RMA$, and $RACT$) to the $WAS$ were 0.46547, 0.56216, and 0.65093, respectively. It implied, taking the supply of other church services per capita ($ROTH$) for instance, that one unit higher for the $ROTH$, on average in 2006, 0.47 unit higher for the $WAS$, if other church services supplies were fixed.

A benefit of using the average value of $PAS$ was that, the $t$-values of these marginal effects could be obtained simply by making some adjustments for the model (Woodbridge, 2002). That was replacing the interaction terms, say ($PAS*ROTH$), by ($PAS-PAS_{Average})*ROTH$, where $PAS_{Average}$ denoted the average value of the $PAS$. After this adjustment, the explanation for the coefficient of the $ROTH$ became the partial contribution of $ROTH$ to $WAS$ when $PAS$ equaled to its average value. Therefore, the $t$-value measured the statistical significance of the estimation for this coefficient. In this study, $t$-values were 8.50, 2.64, and 4.92 respectively for the marginal effects of $ROTH$, $RMA$, and $RACT$, which strongly supported the hypothesis that none of these effects was zero.

Another way of considering $PAS$ was to use its local value for each observation. Since there were spatial variations for the value of $PAS$, the corresponding marginal effects spatially varied in the study area also. Therefore the partial contributions of church services supplies could be mapped to study their spatial patterns. However one thing
should be noticed that, for in the model all variations of the marginal effects linearly derived from the changes of PAS, the spatial patterns of the marginal effects of different church services supplies were perfectly correlated. Figure 5 reported the distribution of the marginal effect of ROTH, RMA, RACT and the value of PAS in space. For this reason, and using the same way of grouping, they presented same spatial patterns. It was that the spatial patterns of the PAS that determined the spatial patterns of these marginal effects.

If this model describes the relationships between the church services supplies and the pastorat performance sufficiently well, the parameter estimations could provide suggestions for pastorat operations. For example at the pastorats with lower PAS, as Uppsala, Gävle, and Hudiksvall, ROTH and RACT contributed more to the WAS, and therefore the pastorat operations should focus more on increasing the numbers of ROTH and RACT. Whilefor the pastorats with higher PAS, such as Skogs, Ockelbo, Nora, or Väddö, RMA had larger effects on the WAS than ROTH and RACT, so their church operations should aim at increasing the number of RMA.
5 Discussions

For a regression model, its practical value is determined by how well it describes the relationship of concern. The methodology used in this study was analyzing the previous models, finding out their similarities and differences to this study, suggesting possible models and modeling strategies, evaluating them with the data available, and finally obtaining the most suitable estimations. This first step of this methodology requires extensive review of previous models. However, for obstacles in language, only materials in English were carefully examined in this study. It could, of course, make several models uncovered during the model scanning, which might increase the possibility of omitting variables. In this study, methods such as Y-lag and FD were used to reduce this possibility.

In addition, during the study of the regression techniques, OLS and GWR were compared to test if the relationship of modeling had any spatial non-stationarity. The comparison was carried out in a form that examining the OLS estimations of the proposed models at first, and then performed the GWR estimation. During the comparing of OLS and GWR, the results of the statistical tests in OLS estimations were used.

However, this way of comparison included a potential problem that might mislead this comparison. In this study, a coefficient of a global model averagely measured the marginal effect of its corresponding variable. However, in some situations, there could be an independent variable that surely affected the dependent variable but this effect was largely different in space. It could be positive in some places of the studying area while be negative elsewhere. If only OLS was employed, this effect was possibly to be averaged, or become statistically insignificant. If it was the case, then using the results from statistical tests for the global model might caused the wrong judgment. However, at present this type of effects was difficult to detect, since the GWR lacked tools for such statistical tests. Besides, for a comparatively small studying area, the possibility of appearing this large and significant spatial variation was tiny. So in this study, the results from comparisons of the OLS first and then GWR were accepted.

Another issue was the discrete zone problem, which means that the estimations at the neighboring locations may be quite different. This difference is not from the uncontinuous spatial process but from the different data determined by the zoning. The GWR could release this problem within the studying area, since the boundaries of zones
are determined endogenously in the process of GWR (Fotheringham et al., 2002). However, in this study, the boundary of the Uppsala diocese was determined exogenously. Therefore, the discrete zone problem could still remain near the boundary of the studying area. One solution for this problem was to include more observations that not only cover the Uppsala diocese but also the regions around it. This approach may be tested by future studies.

Then carried out the GWR and extracted the results for only the Uppsala diocese. However, this approach can only be tested by studies in future.

The $F_1$ and $F_3$ statistics (Leung et al., 2000a,b; Huang and Leung, 2002) were used in this study to test the significances of the spatial variations. However there was a conflict in their results. Specifically, the result of $F_1$ test supported no significant difference between the OLS and GWR estimations, whereas the result of $F_3$ test supported that the coefficient of ROTH, whose estimation was statistically significant in OLS, significantly varied across the studying area. No technical discussion was intended to be performed here. However, a guess for this problem could be that, there was no logical relationship that, if the result of $F_1$ test supported its $H_0$, then the result of $F_3$ test should supported its $H_0$. In this study, the discussion of the partial effects of ROTH didn't reject this guess. However more works were needed to find the strict answer.

The largest limitation for this study was that, no out-of-sample test was performed. All discussions carried out in this study were only based on the in-sample tests. Granger (2005) indicates that, the most important way to evaluate an empirical model, no matter what modeling strategies were applied, is the out-of-sample evaluation. For these models which contain time information, just as models in this study, the evaluation works should be conducted using the data generated after the modeling. For the limitations of the data, this study didn’t perform this test. However, when the data become available, it is necessary to carry out the out-of-sample evaluation.
6 Conclusions

The aim of this study was to develop a multiple linear regression model that measured the relationship between the supplies of church services and the pastorat performance in the Uppsala diocese. By reviewing previous models and examining the nature of data available, several possible models were proposed and parameters of these models were estimated by the OLS. Besides, the geographically weighted regression was used to examine the spatial variations in relationships estimated by these modeling strategies. After several comparisons, the conclusions of this study can be summarized as follows:

- The simple regression and Y-lag models were suffering from the problems of omitted variables. There was no evidence that the FD model had biased parameter estimates.
- No statistically significant spatial variation was detected for the FD model.
- Therefore, by using the ordinary least square estimation for the first-differenced equation, the most suitable model describing the relationship of concern could be obtained. There was no need to concern about the issue of spatial non-stationarity.
- However, the analysis was only based on the in-sample test. In order to obtain more reliable results, out-of-sample test should be performed when more data become available.
Acknowledgement

This thesis is submitted to the Department of Technology and Built Environment at the University of Gävle as part of the MSc. Degree.

Special thanks to Prof. Anders Östman of the University of Gävle. He has provided large amount of help and suggestion to me in the whole procedure of this thesis. Same thanks to Dr. Anders Brandt of the University of Gävle. He has helped me a lot from the fundamental GIS knowledge to the scientific theory and writing.

I thank Mr. Jim Frölander of the GeoIntel System AB. He provided this study topic to me with great supports. I am also grateful to Mr. Johan Lindström for his help with the software and technical matters during the thesis.
Reference


Anselin, L., 2005. *GeoDa 0.9.5i(beta).* [www.geoda.uiuc.edu](http://www.geoda.uiuc.edu), accessed and downloaded on June 1, 2008.


