SPLIT SELECTION METHODS
FOR REGRESSION TREE
ON DETECTING REGIONAL
ECONOMIC CONVERGENCE

Hsin-Ju Chang, Yu-Shan Shih and Tsung-Jen Su
Department of Mathematics, National Chung Cheng University

ABSTRACT

A split selection method for building regression tree in order to identify economic convergence clubs is proposed. It selects the split variable by analyzing residual patterns of some spatial models which are fitted to the data in each node. Through simulations, we find that this method is unbiased and it has more power in selecting informative variables than the method of Postiglione et al (2010). Our split point selection method is more effective than its counterpart as well. Furthermore, the method is applied to a NUTS 2 level data set and some economic clubs are detected.

Key words and phrases: convergence clubs, regression tree, residual analysis, selection bias.

JEL classification: C21, C81.
1. INTRODUCTION

Spatial data which are geographically correlated appear frequently in astronomy, environment, forestry, regional economics, and so forth (Anselin, 1988; Cressie, 1995). Various statistical methods emerge to analyze such data and regression tree is among them. A toy example is given in Figure 1 to illustrate the methodology of classical regression trees. Suppose a data set of size 200 which contains a continuous response variable \( Y \) and two covariates \( X_1 \) and \( X_2 \) is collected and we want to predict \( Y \) based on the covariates. Starting at the root node which contains all the data, classical regression trees recursively partition the data in each node into two children nodes. In this example, the tree first splits the data based on the value of \( X_1 \). If the value is less than or equal to 5, its observation goes to the left node; otherwise the observation goes to the right node. Among those 160 observations, it further splits the data based on the value of \( X_2 \) which is a categorical variable. This data splitting procedure stops based on some criteria. Afterward, a model is fitted and the predicted values for \( Y \) are obtained in each terminal node. For example, a constant function is fitted and the predicted values are 25 for all 40 cases in node 1. Similarly, the predicted values are given in node 2 and 3. For future observations, the prediction can be made based on the resulting tree. Thus, a regression tree can be treated as a piecewise approximation to the real response surface.

Fig. 1: A toy example of regression tree. The split beside each intermediate node channels each case to the left node, if it is satisfied; otherwise to the right. The numbers beside each terminal node are the sample size and the corresponding estimate of the responses.
Three basic elements including split selection, stopping rule and node fitting are involved in tree construction procedure. Breiman et al (1984) gives a complete description of the methodology. Regression tree has the reputation of easy interpretation (Breiman et al, 1984; Loh, 2002). Some comparison studies show that it has good prediction power as well (Kim et al, 2007; Loh, 2008). Recently, effort has been made on using regression tree methods to identify convergence clubs in regional economics (Postiglione et al, 2010). This paper mainly focuses on split selection methods for regression tree on detecting regional convergence clubs.

Based on the data, on the one hand, a split on an ordered variable \( X \) with \( N \) distinct values yields \((N - 1)\) splits of the form \( X \leq c \). On the other hand, if \( X \) is a categorical variable with \( L \) categories, there are \((2^L - 1)\) splits of the form \( X \in A \) where \( A \) is a non-empty subset of the \( X \) values. One way to decide the split at each node is by using node “impurity” which is usually defined as the sum of squared errors at that node where a constant or least square regression model is fitted to the data. The split which minimizes the sum of the impurities in the two children nodes is selected. Breiman et al (1984) proposed the CART algorithm in which they used the exhaustive search method to find such split. That is, the minimizer is obtained by searching over all possible \( c \) and \( A \) among all the covariates. However, it is well known that such approach has selection bias (Loh, 2002). It tends to select variables which provide more split points. To overcome this bias problem, Loh (2002) considered the split variable selection and the split point selection separately. Residual pattern analysis of the usual linear regression models is conducted to select the split variable. The split point is then decided by comparing the sums of squared errors. After the split is decided at each node, the data are channeled into two subnodes and the process continues recursively. The final tree is determined by either a direct stopping rule (like the node size is small) or a pruning procedure (Breiman et al, 1984; Loh, 2002).

Postiglione et al (2010) implemented a regression tree method to analyze spatial data in regional economics. To count for spatial dependence, their method fits a spatial model, in stead of a constant or least square model, to the growth rate at each node. The covariates are used only for splitting the nodes and the split is determined by an exhaustive search scheme similar to that of Breiman et al (1984). In this paper,
we show that their split selection method inherits the same selection biases as in the method of Breiman et al (1984). By contrast, based on the approach of Loh (2002), we propose an alternative split variable selection method which takes spatial dependence into consideration when computing the related residuals. This alternative method is shown to be unbiased and is more powerful in selecting informative variables than that of Postiglione et al (2010) in our simulation studies. An effective split point selection method is proposed as well. Furthermore, in a real data study, we apply our method to a NUTS 2 level data set. The resulting tree provides an empirical evidence of regional convergence clubs in the European Union.

The rest of this paper is organized as follows. A brief review of two spatial convergence models in regional economics is given in Section 2. We introduce the split selection method of Postiglione et al (2010) and our alternative method in Section 3. Simulation experiments are conducted and comparisons of the two split variable selection methods and the two split point selection methods are given in Section 4. A real data analysis is given in Section 5. Finally, conclusion and some remarks are given in Section 6.

2. ECONOMIC CONVERGENCE MODELS

In regional economics, Barro and Sala-i Martin (1991, 1995) and Mankiw et al (1992) suggested the following model to measure the convergence in a cross section of economics, for region $i$,

$$
\ln(y_{t,i}/y_{0,i}) = \alpha + \beta \ln(y_{0,i}) + \varepsilon_i,
$$

(1)

where $y_{t,i}$ is the per capita GDP at final time $t$, $y_{0,i}$ is the initial level of per capita GDP, $(y_{t,i}/y_{0,i})$ is the growth rate, $\beta = -(1 - e^{-bt})$, $b$ is the speed of convergence which measures how fast economies will converge toward the steady state and $\varepsilon_i$ are the error terms, which are i.i.d. normally distributed with mean 0 and variance $\sigma^2_\varepsilon$. If the null hypothesis ($\beta = 0$) is rejected in favor of the alternative hypothesis ($\beta < 0$), we can conclude, on an empirical basis, those regional economies converge to the same level of per-capita income (Arbia, 2006, p. 12-13).
In recent decades, empirical evidences supported the idea of convergence clubs: regions within a country or integrated area (such as the European Union) might experience not so much a global convergence process, but instead a convergence by “clubs”, having the initial conditions in common (Arbia, 2006, p. 138).

Many authors have proposed methods to test the hypothesis of convergence clubs and to detect such clubs (Durlauf and Johnson, 1995; Desdoigts, 1999; Corrado et al, 2005; Maasoumi et al, 2007; Postiglione et al, 2010; Bartkowska and Riedl, 2012). Durlauf and Johnson (1995) considered the number of clubs as an endogenous parameter and proposed a regression tree algorithm to identify convergence clubs. However, regional data are usually geographically correlated and model (1) is not capable to catch such spatial dependence (Anselin, 1988). Two models which take spatial correlation among observations (regions) into account are considered. One is the *Spatial Autoregressive Regression* (SAR) model and the other is the *Spatial Error model* (SEM) (see, Arbia (2006) for details). These two models are introduced in the following.

Let $W$ be a weight spatial matrix,

$$ W = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & \cdots & w_{n,n} \end{pmatrix}, $$

where $w_{i,j}$ usually depends on the distance between the centroids of regions $i$ and $j$ (Arbia, 2006).

The SAR model is

$$ h(\ln(y_{t,i}/y_{0,i})) = \alpha + \beta \ln(y_{0,i}) + \omega_i, \quad (2) $$

where

$$ h(\ln(y_{t,i}/y_{0,i})) = \ln(y_{t,i}/y_{0,i}) - \rho \sum_j w_{ij} \ln(y_{t,j}/y_{0,j}), $$

$\rho$ is a scalar spatial autoregressive parameter and $\omega_i$ are i.i.d. normally distributed. This model takes spatial dependence into account by including an autoregressive term of the dependent variable.
An alternative to the SAR model is the following SEM model where spatial dependence appears in both the independent and the dependent variables.

\[ k(\ln(y_{t,i}/y_{0,i})) = \alpha + \beta[k(\ln(y_{0,i}))] + u_i, \]  

(3)

where

\[ k(\ln(y_{t,i}/y_{0,i})) = \ln(y_{t,i}/y_{0,i}) - \delta \sum_j w_{ij} \ln(y_{t,j}/y_{0,j}), \]

\[ k(\ln(y_{0,i})) = \ln(y_{0,i}) - \delta \sum_j w_{ij} \ln(y_{0,j}), \]

\( \delta \) is a scalar parameter on the spatially correlated errors and \( u_i \) are i.i.d. normally distributed. Either the SAR or the SEM model only involves the GDP variable. No covariates are included in the models.

Instead of fitting a single SAR or SEM model, Postiglione et al (2010) incorporated some related covariates and proposed piecewise SAR or SEM models using regression tree. They modified the classical regression tree procedure (Breiman et al, 1984; Durlauf and Johnson, 1995) to accommodate the spatial dependence among economic regions by fitting a SAR or SEM model at each node. The method uses covariates only for data splitting which is described in the following section. Basically, they used spatial regression tree methods to exploit the alternative concept of convergence clubs. In this paper, we follow their identification approach, but we have an alternative viewpoint toward split selection.

3. SPLIT SELECTION METHODS

Split variable and point selection methods for constructing spatial regression trees are discussed in this section. First, the selection scheme proposed by Postiglione et al (2010) which uses the exhaustive search method of Breiman et al (1984) is introduced. This selection method considers split variable and point selection simultaneously. However, it is known that the method of Breiman et al (1984) has selection bias toward variables with more split points when it is used in classical regression trees (Loh, 2002). Therefore, it is reasonable to believe that the method used in Postiglione et al (2010) has selection bias as well. We then propose a new method which separates variable
selection from point selection by following the idea of Loh (2002). Residual analysis is
carried out to find the split variable and point. In the selection methods, the covariates
are only used to split nodes. They are not included in fitting either the SAR or the
SEM model at each node.

3.1. Exhaustive search method

Postiglione et al (2010) finds the split which yields the maximum distance between
the parameters of the estimated models. For each covariate $X$, a split of the form
$X \leq c$, if $X$ is an ordered variable; or $X \in A$, if $X$ is a categorical variable is considered.
Suppose at node $\tau$, such a split creates the subnodes $\tau_l$ and $\tau_r$. Let $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ be the
MLE of parameter vector $(\alpha, \beta)$ in the SAR (Eq. (2)) or SEM (Eq. (3)) model. Denote
$\hat{\theta}_{\tau_l}$ and $\hat{\theta}_{\tau_r}$ be the MLEs for the data in the subnode $\tau_l$ and $\tau_r$ respectively. Similarly,
$\Sigma_{\tau_l}$ and $\Sigma_{\tau_r}$ are their estimated covariance matrices respectively. By assuming that $\hat{\theta}_{\tau_l}$
and $\hat{\theta}_{\tau_r}$ are independent, Postiglione et al (2010, p. 2778) asserts that the statistic

$$I_X(\tau) = (\hat{\theta}_{\tau_l} - \hat{\theta}_{\tau_r})^T(\Sigma_{\tau_l} + \Sigma_{\tau_r})^{-1}(\hat{\theta}_{\tau_l} - \hat{\theta}_{\tau_r})$$

(4)

follows an asymptotic chi-squared distribution with 2 degrees of freedom. All possible
constants $c$ or subsets $A$ and covariate $X_k$, $k = 1, \ldots, p$ are considered and the variable
associated with the maximum value of the $I_{X_k}(\tau), k = 1, \ldots, p$ is the split variable at
node $\tau$. We denote this method as the ES (Exhaustive Search) method.

3.2. Residual pattern analysis

Recall that an ordered $X$ variable with $N$ distinct values creates $(N - 1)$ splits.
Suppose $X_1$ and $X_2$ are two such variables with $N_1$ and $N_2$ distinct values, respectively.
If $N_1 > N_2$ and both variables are independent of the response variable, then $X_1$
has a higher chance to be selected than $X_2$. The situation is the same for categorical
covariates. Thus, variable selection bias appears and it is caused by the fact that the
split variable and the split point are considered simultaneously. To avoid such bias, Loh
(2002) treats split variable selection and split point selection separately. In the process
of selecting split variables, Loh (2002) analyzes the associated residual pattern of each covariate or interaction of two covariates. Later, the \( p \)-values of the corresponding Pearson chi-squared tests are used to select split variables for classical regression trees. The method itself is able to eliminate variable selection bias and its success relies on its residual pattern analysis. We follow the same idea to select the split variables for spatial regression trees in our setting. After fitting spatial SAR or SEM model to the data in each node, we compute the associated residuals. The variable corresponding to the most significant residual pattern which is indicated by the \( p \)-value of the corresponding Pearson’s chi-squared test is selected. We denote our method as the RA (Residual Analysis) method and the algorithm is given in the following. The steps follow exactly those in Algorithm 1 and 2 of Loh (2002) except that residuals are obtained by fitting the spatial regression models to the data in each node. Furthermore, the algorithm is capable of selecting the informative single covariate as the split variable whether it acts as a main factor (step 2 and 3) or it interacts with the other covariate (step 4 to 6). If the interaction effect is stronger than any main effect and both covariates are numerical, the variable with the smaller total SSE is selected (step 7(a)). Otherwise, the categorical variable with the smaller \( p \)-value is used as the split variable (step 7(b)).

**Algorithm**

1. Obtain the residuals from a SAR or SEM model fitted to the response in each node.

2. For each numerical-valued variable, divide the data into four groups at the sample quartiles; construct a \( 2 \times 4 \) contingency table with the signs of the residuals (positive versus non-positive) as rows and the groups as columns; count the number of observations in each cell and compute the \( \chi^2 \)-statistic and its theoretical \( p \)-value.

3. Do the same computation for each categorical variable, using the categories of the variable to form the columns of the contingency table and omitting columns with zero column totals.

4. To detect interactions between a pair of numerical-valued variables \((X_k, X_l)\), divide the \((X_k, X_l)\)-space into four quadrants by splitting the range of each variable.
into two halves at the sample median; construct a $2 \times 4$ contingency table using
the residual signs as rows and the quadrants as columns; compute the $p$-value. Again, columns with zero column totals are omitted. We refer to this as an interaction test.

5. Do the same computation for each pair of categorical variables, using their value pairs to divide the sample space. For example, if $X_k$ and $X_l$ take $c_k$ and $c_l$ values, respectively, the $p$-value are computed from a table with two rows and number of columns equal to $c_k c_l$ less the number of columns with zero totals.

6. For each pair of variables $(X_k, X_l)$ where $X_k$ is numerical-valued and $X_l$ is categorical, divide the $X_k$-space into two at the sample median and the $X_l$ space into as many sets as the number of categories in its range (if $X_l$ has $c_\lambda$ categories, this splits the $(X_k, X_l)$-space into $2 c_\lambda$ subsets); construct a $2 \times 2 c_\lambda$ contingency table with the signs of the residuals as rows and the subsets as columns; compute $p$-value for the table after omitting columns with zero totals.

7. If the smallest $p$-value is from step 2 to 3, select the associated $X$ variable to split the node. Otherwise,

(a) If both variables are numerical-valued, the node is split in turn along the sample mean of each variable; for each split, the total sum of the squared errors (SSE) for a SAR or SEM model is obtained for each subnode; the variable yielding the split with the smaller total SSE is selected.

(b) Otherwise if at least one variable is categorical, the one with the smaller $p$-value from step 2 to 3 is selected.
3.3. Point selection methods

After the split variable is selected, the split point needs to be determined. Let $X$ denote the selected split variable. We study two point selection methods. The impurity criterion of Equation (4) is a naive choice. The best split point is the one which maximizes $I_X(\tau)$ for node $\tau$. This is the method proposed by Postiglione et al (2010) and we denote it as the $I$ method. Another approach is to compare residuals. For every possible split point, two subsets of the data at current node $\tau$ are created. By fitting a SAR or SEM model which does not contain any covariates to each subset, we can obtain its corresponding sums of squared errors, $SSE_{\tau_l}$ and $SSE_{\tau_r}$. The chosen point is the one which minimizes $SSE_{\tau_l} + SSE_{\tau_r}$. This approach was used in Breiman et al (1984) and Loh (2002) for classical regression trees. We denote it as the $R$ method.

4. SIMULATION EXPERIMENTS

We conduct some simulation experiments to compare the split variable and point selection methods. First, in Section 4.1, two variable selection methods, ES and RA, are compared when the response variable is independent of the covariates. Next, in Section 4.2, some models where the response depends on some covariates are simulated and we study the variable selection power of the ES and RA methods applied to these models. Finally, in Section 4.3, the two point selection methods, $I$ and $R$, are compared on some models with change points.

The initial level of per capital GDP, $y_0$, is generated from an uniform distribution on $(0, 1000)$. The coordinate of each regional centroid is uniformly distributed on $(0, 10) \times (0, 10)$. The values for parameters $\alpha$, $\beta$, $\rho$ and $\delta$ are chosen as the corresponding values in Postiglione et al (2010, p. 2781-2783). The $(i,j)$ entry of the weight spatial matrix, $W$, is defined as $w_{i,j} = e^{-d_{i,j}} / \sum_j e^{-d_{i,j}}$ where $d_{i,j}$ is the distance between the centroids of regions $i$ and $j$. This means that $W$ is row standardized. The error terms $\omega_i$ and $u_i$ are normal distributed with mean 0 and standard deviation 4. Two hundred observations are generated from the SAR and the SEM models. The frequency of each covariate being selected by the ES and the RA methods is counted and the simulation
is repeated five hundred times. All simulations are performed in R (R Development Core Team, 2012) and package spdep (Bivand R. et al., 2011) is used to fit the SAR and SEM models and to obtain the related parameter estimates and residuals.

Table 1: Distributions of $X$ variables used in the simulation studies. $Z$, $E$, $U_{10}$, $C_2$, and $C_{10}$ are mutually independent; $Z$ is a standard normal variable; $E$ is an exponential variable with mean one; $U_{10}$ is a uniformly distributed variable on integer values of [0, 10]; $C_m$ denotes a $m-$category variable taking values \{1, 2, ..., $m$\} with equal probabilities.

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z$</td>
<td>$E$</td>
<td>$U_{10}$</td>
<td>$C_2$</td>
<td>$C_{10}$</td>
</tr>
</tbody>
</table>

4.1. Independent Models

In this study, the response variable $Y$ is designed to be independent of the $X$ covariates. In this setting, a good split variable selection method should be able to select each covariate with equal probability. A selection method with such property is named an unbiased method.

The distributions of the $X$ covariates are given in Table 1 while the distribution of $Y$ follows a SAR or SEM model. The parameter values are $\alpha = 0.611, \beta = -0.127$ and $\rho = 0.818$ for the SAR model (Eq. (2)) and $\alpha = 0.241, \beta = -0.121$ and $\delta = 0.842$ for the SEM model (Eq. (3)) which are the corresponding values in Postiglione et al (2010, p. 2781-2783). The relative frequency of each covariate selected by the ES and RA methods is shown in Figure 2.

From Figure 2, we find that, for the three numerical covariates, the ES method selects $X_1$ or $X_2$ more frequently than $X_3$ under either the SAR or SEM model. It is observed that $X_3$ has less split points than that of $X_1$ or $X_2$ (10 vs. 199). Similarly, $X_4$ has less split points than that of $X_5$. The ES method selects $X_4$ less often than $X_5$. These facts indicate that the ES method has selection bias. It tends to select variables
with more split points. On the other hand, the RA method selects each covariate all within three standard errors of 0.2 under either the SAR or SEM model. Thus, the RA method is unbiased.

Fig. 2: Estimated probabilities of variable selection for the RA and the ES method where \( Y \) is independent of the \( X \)'s. The observations of \( Y \) are generated either from a SAR or SEM model. The parameter values are \( \alpha = 0.611, \beta = -0.127 \) and \( \rho = 0.818 \) for the SAR model and \( \alpha = 0.241, \beta = -0.121 \) and \( \delta = 0.842 \) for the SEM model. The distributions of \( X \)'s are given in Table 1. The simulation standard errors are about 0.018.

4.2. Dependent Models

In this study, we compare two variable selection methods when the response variable depends on some covariates. A good selection method should be able to choose informative covariate(s) with higher probabilities. Six designs are used. The first four designs (A, B, C, and D in Table 5) present models where the slope term affects the response. For models in Design A and B, \( X_3 \) is the only informative covariate. For models in Design C and D, both \( X_3 \) and \( X_4 \) are informative. Furthermore, interaction between \( X_3 \) and \( X_4 \) is the only informative term in Design D. Since the selection methods choose only single covariate as the split variable, either \( X_3 \) or \( X_4 \) is a valid split variable in these settings. The last two designs (I and II in Table 6) give models where the intercept term including \( X_3 \) alone affects the response. The ES and RA methods
are compared on these simulated models and the estimated probability of informative covariate(s) being selected by the two methods are recorded. The results for Design A and B are given in Figure 5 while those for Design C and D are shown in Figure 6. Similarly, the results for Design I and II are shown in Figure 7 respectively.

Figure 5 and 6 show that as parameter $\beta$ increases, the selection probability for either the RA or the SE method increases except in some cases. Since the value of $\beta$ indicates the strength of the informative variable(s), the outcome of increasing selection probability is expected. For those lines do not appear to be monotone increasing (like the lines for the ES method on the left panel of Figure 5), they are the cause of simulation errors. Furthermore, we find that the RA method always has higher probability of selecting the informative variable(s) than that of the ES method under the spatial design models. Moreover, in some cases, like Design D, the selection probability for the ES method is even less than $1/5 = 0.2$ which is the probability of random selection. Actually, the selection probability for the ES method is less than 0.2 at some points across all designs. Similar results also appear in Figure 7 where $\alpha$ value is changed.

4.3. Split point study

Two sets of designs are used to investigate the two split selection methods. They are listed in Table 7 and the models are used to generate 200 observations. For each simulated data set, the I and R methods are used to select the split point. The point value is recorded and the simulation is repeated 500 times. The distributions of split points for the two methods are shown in Figure 8, 9, 10 and 11.

For the first set of designs, the best split point is the associated median. The closer to the median the split point is, the better the split selection method is. From Figure 8 and 9, we observe that the R method is better than the I method for all $\beta$ values. Moreover, as the $\beta$ value increases, the distribution for the R method becomes more concentrated. For the other set of designs, the best split points are at one standard deviation from the associated mean. Thus, there are two such points and both Figure 10 and 11 show that the distributions for either the I method or the R method are bi-
modal. As the $\beta$ value increases, the distributions become more concentrated. Again, we find that the R method performs better than the I method for all $\beta$ values.

Table 2: Regions under study

<table>
<thead>
<tr>
<th>Country</th>
<th>Austria</th>
<th>Belgium</th>
<th>Finland</th>
<th>France</th>
<th>Germany</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regions</td>
<td>9</td>
<td>11</td>
<td>5</td>
<td>22</td>
<td>38</td>
</tr>
<tr>
<td>Country</td>
<td>Greece</td>
<td>Ireland</td>
<td>Netherlands</td>
<td>Portugal</td>
<td>Sweden</td>
</tr>
<tr>
<td>Regions</td>
<td>13</td>
<td>2</td>
<td>12</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

The distribution of the split point found by the I method has higher variation than that of the R method across all designs. This outcome may be caused by the assumption of independence between two parameter vector estimators $\hat{\tau}_l$ and $\hat{\tau}_r$ used in Equation (4). Since the two estimators are obtained from modeling two correlated subsets of data, the assumption is probably invalid. As a result, the split point chosen by the I method has higher variability.

In summary, we find that the split selection method of Postiglione et al (2010) attends to select variables with more split points. By separating variable selection from split point selection, our method based on residual pattern analysis is able to eliminate such bias. Moreover, it has higher chance of selecting the informative covariates than its counterpart when the responses depend on the covariates in our simulation study. The split point selected by our method is less variable than that of Postiglione et al (2010).

We have studied another weight spatial matrix, $w_{i,j} = v_{i,j} / \sum_j v_{i,j}$ where $v_{i,j} = I\{0 < d_{i,j} < u\}$ which indicates neighbor regions of region $(i, j)$. The simulation results, where $u = 5$ is used, are similar to what you have obtained in this section. The weight spatial matrix $W$ may have effect on the SAR or the SEM models (Arbia, 2006). However, since our split selection method relies on residuals, the effect of the matrix may not be as strong as it may have on the models at each node.
5. EMPIRICAL STUDY

A NUTS 2 level data set collected from the Eurostat website http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/ is studied. It consists of 127 regions of Europe listed in Table 2. Per capita GDP from 1997 to 2005, measured in purchasing power standard, are used to obtain the growth rate for each region. The standardized covariates for year 2005 are listed in Table 3. The covariates here are used for node splitting only and they are not used in fitting models at each node. The weight spatial matrix, $W$, like what we used in the simulation study, is $w_{i,j} = e^{-d_{i,j}} / \sum_j e^{-d_{i,j}}$ where $d_{i,j}$ is the distance between the centroids of regions $i$ and $j$. We apply our split selection method to investigate possible convergence clubs in these regions.

We first check spatial dependence within the data using some standard diagnostics tests (Anselin et al, 1996) and the results are given in Table 4. The significant $p$-values suggest spatial dependence among these regions. We then recursively partition the data using our RA method to select the split variable and the R method to select the split point. At each node, a SEM regression model is fitted. A direct stopping rule is used to terminate the splitting process. We stop splitting the data if the size of a node is less than a threshold (25 in this study) or the $p$-value of the slope ($\beta$) estimate is greater than 0.05. The resulting tree is given in Figure 3 and the corresponding regional map is given in Figure 4. Besides, a SAR model is fitted at each node and a tree is obtained by the same stopping rule. However, it yields 4 terminal nodes with only one (convergence) club including 37 regions. This tree gives more terminal nodes and less economic clubs than the SEM tree. Therefore, we only report the SEM tree.
The tree shows that based upon the standardized values of Investment in services, the regions are divided into three nodes (clubs). The regions with values of Investment in services above 0.245 are in node 3. If the corresponding values are below or equal to -0.52, regions are in node 1. The remaining regions are in node 2. We find that among 127 regions, 86 of them form a convergence club (node 2) with estimate $\hat{\beta} = -0.013$ ($p$-value = 0.098). On the other hand, 21 regions (node 3) show economic divergence ($\hat{\beta} = 0.049$ with $p$-value < $10^{-2}$) and 20 regions (node 1) show no sign of either economic convergence or divergence. Our resulting clubs are not conterminous. Similar results were also found by Postiglione et al (2010) who asserts that regions which are not contiguous can form convergence clubs. The convergence club includes most regions of Austria (8), Belgium (10), Finland (4), Netherlands (10), Portugal (6), and Sweden (7) where the number of regions is in parentheses. Some regions of France (13), Germany (20), and Greece (8) are also in this club. On the other hand, the divergence club consists of some regions of France (6), Germany (7), Ireland (2) and Netherlands (2). One region of Belgium, Finland, Greece and Sweden are also in this club. In summary, we are able to identify one convergence and one divergence club by applying our proposed method to this NUTS 2 data set.
Investment in services ≤ −0.520

\[ n = 20 \]
\[ \hat{\alpha} = 0.134 \]
\[ \hat{\beta} = 0.016 \]

Investment in services ≤ 0.245

\[ n = 86 \]
\[ \hat{\alpha} = 0.461^* \]
\[ \hat{\beta} = -0.013^* \]

\[ n = 21 \]
\[ \hat{\alpha} = 0.923^* \]
\[ \hat{\beta} = 0.049^* \]

Fig. 3: Tree-structured SEM model for the NUTS data. The split beside each intermediate node channels each case to the left node, if it is satisfied; otherwise to the right. The numbers beside each terminal node are the sample size and its corresponding parameter estimates of the SEM model (Eq. (3)). An asterion indicates the estimate is significant at 10% level.

Fig. 4: Maps of economic regions using the tree-structured SEM model.
6. CONCLUSION AND REMARKS

In this paper, we introduce a new split selection method for constructing spatial regression trees in order to identify economic convergence clubs. By considering variable selection and point selection separately, we provide an unbiased split selection method. After fitting a SAR or a SEM model at each node, residual pattern analysis is conducted and Pearson’s chi-squared tests are performed to obtain the split variables. The split point is chosen by comparing the sum of the sums of squared errors at two subnodes. Through simulations, we show that our method is unbiased and is more powerful in selecting the informative variables than that of Postiglione et al (2010). Furthermore, the split point selected by our method is less variable than its counterpart. We apply our method to analyze NUTS data where an economic convergence club and an economic divergence club are identified. This empirical finding may contribute to the study of regional economic convergence.

In the data analysis, we use 0.10 as a threshold for multiple testings to determine the significance of the slope estimators. The threshold should depend on the tree structure itself. A better way to make statistical inference at each terminal node is needed. Some variables such as social capital may be relevant for identifying convergence clubs. But we was not able to collect them. Also, in our empirical study, a SEM model is fitted at each node. It is certainly feasible to fit other spatial models like those in Arbia (2006), in the tree building process. Furthermore, the resulting tree is selected by a direct stopping rule. A better way to select the best subtree is to use the pruning method of Breiman et al (1984). These are the subjects of our future research.

Acknowledgements

The authors would like to thank Prof. Pei-Sheng Lin and Prof. Jyh-Lin Wu for their helpful comments and suggestions. The authors are grateful to the Editor, the Associate Editor and the referee for the comments which improve the presentation of this article.
Table 5: Models for power studies of the variable selection methods where the slope term of the SAR or the SEM model depends on some functions of the covariates. The distributions of \( X \) are given in Table 1. Functions used in the slope term \( f_1(X_3) = 1 - 0.5 \times I(X_3 < 5), f_2(X_3) = 1 - 0.5 \times I(|X_3 - 5| > 2), f_3(X_3, X_4) = 1 - 0.5 \times I(X_3 < 5 \text{ or } X_4 = \{2\}), \) and \( f_4(X_3, X_4) = 1 - 0.5 \times I(X_3 < 5 \text{ and } X_4 = \{2\}) \). Other parameter values are \( \rho = 0.818, \quad \delta = 0.842 \quad \text{and} \quad \beta = 0.05, 0.10, 0.15, \text{or} \quad 0.20. \)

<table>
<thead>
<tr>
<th>Design</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>SAR: ( h(\ln(y_{1,i}/y_{0,i})) = 0.611 - \beta f_1(X_3) \ln(y_{0,i}) + \omega_i ) &lt;br&gt; SEM: ( k(\ln(y_{1,i}/y_{0,i})) = 0.241 - \beta f_1(X_3) k(\ln(y_{0,i})) + u_i )</td>
</tr>
<tr>
<td>B</td>
<td>SAR: ( h(\ln(y_{1,i}/y_{0,i})) = 0.611 - \beta f_2(X_3) \ln(y_{0,i}) + \omega_i ) &lt;br&gt; SEM: ( k(\ln(y_{1,i}/y_{0,i})) = 0.241 - \beta f_2(X_3) k(\ln(y_{0,i})) + u_i )</td>
</tr>
<tr>
<td>C</td>
<td>SAR: ( h(\ln(y_{1,i}/y_{0,i})) = 0.611 - \beta f_3(X_3, X_4) \ln(y_{0,i}) + \omega_i ) &lt;br&gt; SEM: ( k(\ln(y_{1,i}/y_{0,i})) = 0.241 - \beta f_3(X_3, X_4) k(\ln(y_{0,i})) + u_i )</td>
</tr>
<tr>
<td>D</td>
<td>SAR: ( h(\ln(y_{1,i}/y_{0,i})) = 0.611 - \beta f_4(X_3, X_4) \ln(y_{0,i}) + \omega_i ) &lt;br&gt; SEM: ( k(\ln(y_{1,i}/y_{0,i})) = 0.241 - \beta f_4(X_3, X_4) k(\ln(y_{0,i})) + u_i )</td>
</tr>
</tbody>
</table>

Table 6: Models for power studies of the variable selection methods where the intercept term of the SAR or the SEM model depends on some functions of the covariates. The distributions of \( X \) are given in Table 1. Functions used in the intercept term are \( f_1(X_3) = 1 - 0.5 \times I(X_3 < 5), \) and \( f_2(X_3) = 1 - 0.5 \times I(|X_3 - 5| > 2) \). Other parameter values are \( \rho = 0.818, \quad \delta = 0.842, \quad \alpha = 0.5, 1.0, 1.5, \text{or} \quad 2.0. \)

<table>
<thead>
<tr>
<th>Design</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>SAR: ( h(\ln(y_{1,i}/y_{0,i})) = 0.60 \times f_1(X_3) - 0.2 \times \ln(y_{0,i}) + \omega_i ) &lt;br&gt; SEM: ( k(\ln(y_{1,i}/y_{0,i})) = 0.2a \times f_1(X_3) - 0.2 \times k(\ln(y_{0,i})) + u_i )</td>
</tr>
<tr>
<td>II</td>
<td>SAR: ( h(\ln(y_{1,i}/y_{0,i})) = 0.60 \times f_2(X_3) - 0.2 \times \ln(y_{0,i}) + \omega_i ) &lt;br&gt; SEM: ( k(\ln(y_{1,i}/y_{0,i})) = 0.2a \times f_2(X_3) - 0.2 \times k(\ln(y_{0,i})) + u_i )</td>
</tr>
</tbody>
</table>

Table 7: Models for split point studies of the point selection methods where the slope term of the SAR or the SEM model depends on some functions of the covariates. The distributions of \( X \) are given in Table 1. Functions used in the slope term are \( f_{s_1}(X_l) = 1 - 0.5 \times I(X_l < q_{l,5}) \) where \( q_{l,5} \) is the median of \( X_l \) and \( f_{s_2}(X_l) = 1 - 0.5 \times I(|X_l - \mu_l| > \sigma_l) \) where \( \mu_l \) and \( \sigma_l \) are the mean and standard deviation of \( X_l, \ l = 1, 2, 3. \) Other parameter values are \( \rho = 0.818, \quad \delta = 0.842, \quad \beta = 0.05, 0.10, 0.15, \text{or} \quad 0.20. \)

<table>
<thead>
<tr>
<th>Design</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_{1t} )</td>
<td>SAR: ( h(\ln(y_{1,i}/y_{0,i})) = 0.611 - \beta f_{s_1}(X_l) \ln(y_{0,i}) + \omega_i ) &lt;br&gt; SEM: ( k(\ln(y_{1,i}/y_{0,i})) = 0.241 - \beta f_{s_1}(X_l) k(\ln(y_{0,i})) + u_i )</td>
</tr>
<tr>
<td>( S_{2t} )</td>
<td>SAR: ( h(\ln(y_{1,i}/y_{0,i})) = 0.611 - \beta f_{s_2}(X_l) \ln(y_{0,i}) + \omega_i ) &lt;br&gt; SEM: ( k(\ln(y_{1,i}/y_{0,i})) = 0.241 - \beta f_{s_2}(X_l) k(\ln(y_{0,i})) + u_i )</td>
</tr>
</tbody>
</table>
Fig. 5: The estimated probability of $X_3$ being selected for the RA and the ES method under Design A and B.
Fig. 6: The estimated probability of $X_3$ or $X_4$ being selected for the RA and the ES method under Design C and D.
Fig. 7: The estimated probability of $X_3$ being selected for the RA and the ES method under Design I and II.
Fig. 8: Distributions of split points for the point selection methods where the models are generated by the SAR model under Design $S_{11}$, $S_{12}$ and $S_{13}$. Solid curve is for the R method; dashed curve for the I method. The vertical line represents the location of the correct split point.
Fig. 9: Distributions of split points for the point selection methods where the models are generated by the SEM model under Design $S_{11}$, $S_{12}$ and $S_{13}$. Solid curve is for the R method; dashed curve for the I method. The vertical line represents the location of the correct split point.
Fig. 10: Distributions of split points for the point selection methods where the models are generated by the SAR model under Design $S_{21}$, $S_{22}$ and $S_{23}$. Solid curve is for the R method; dashed curve for the I method. The vertical lines represent the locations of the correct split points.
Fig. 11: Distributions of split points for the point selection methods where the models are generated by the SEM model under Design $S_21$, $S_22$ and $S_23$. Solid curve is for the R method; dashed curve for the I method. The vertical lines represent the locations of the correct split points.
References


[Received August 2013; accepted May 2014.]
用以偵測區域經濟收斂的迴歸樹其分割選取方法

張心如 史玉山 蘇宗仁
國立中正大學數學系

摘要
我們提出迴歸樹中新的分割選取方法來認定經濟體成長率的收斂與否。在選取切割變數上，我們利用了空間模型上殘差的特徵分析。經由模擬實驗得知此方法具有選取的不偏性且較徹底搜尋法更能有效地選取到變數。在切割點的判定上，我們亦提出更有效率的方法。我們利用此迴歸樹方法分析歐盟 NUTS 資料並認定其中部分經濟體成長率的收斂性。

關鍵詞: 迴收斂經濟體, 迴歸樹, 殘差的特徵分析, 選取偏差。

JEL classification: C21, C81.