Support vector machine applied to land use prediction using socio-economic factors in a compact city model

Luis C. Manrique R., Kayoko Yamamoto

***Abstract*—**This paper proposes a method to evaluate the socio-economic factors in a compact city model using Support Vector Machine (SVM). Socio-economic factors are evaluated in the urbanization promotion area of the Aomori metropolitan area in Japan. By using these variables it was possible to predict the land use types using SVM and evaluating the area under the curve (AUC) for each predictor. The results showed that for the residential area the accuracy was higher than 93%. Appropriate Kappa and Rand indexes were found through simulations with values exceeding 0.8. AUC was applied to the predictors and by using this algorithm it was possible to identify the most important socio-economic factors for each class. It is possible to reduce the number of classes according to AUC comparison. Finally, we have contributed to integrate geospatial information with socio-economic factors in order to classify land use.

***Keywords*—**Compact city model, GIS, Socio-economic factors, Support Vector Machine.

I. Introduction and framework

A. Background and purpose

Urban planning has been one of the most important challenges for the humankind. Nowadays there are different kinds of models applied to urban planning such as sustainable, green, smart and compact cities. We will focus on the compact city model because it is one of the most important urban designs applied to cities in recent years; and it mainly address the issues faced by population density, central area revitalization, mixed-use development, services and facilities.

There are different techniques to study the land use prediction, for instance by remote sensing techniques[1], linear regression models, classification trees and others. The purpose of this study is to propose a method for land use classification by analyzing several socio-economic factors in a compact city model. In this study we apply a Support Vector Machine (SVM) to classify the land use in a compact city model through socio-economic factors, we perform simulations using grid search algorithm to calculate the best cost and gamma parameters by tuning the model. Finally we estimate the area under the curve (AUC) to choose the best classifiers. Information is processed through a Geographic Information System (GIS) which combines hardware and software to manipulate, store, retrieve, view and analyze large spatial database[2].

This paper consists of 7 sections. Section II presents the methods for land use classification through SVM evaluation using socio-economic factors. The section will explore the haversine distance calculation, SVM model and AUC analysis. Section III introduces the outline of the study area, here we will present the study area description, it will explain special characteristics of the metropolitan area (MtA) and the land use master plan of Japanese prefectures. Based on these, section IV explains the data processing, including the characteristics of Japanese land use system and socio-economic factors chosen for this analysis. Section V presents the results for the compact city's land use through the SVM. It offers the results of simulations in order to get the most accurate parameters for the model and the AUC results. It will show the comparison between the original and predicted data. Section VI provides discussion on this study, presenting the characteristics of the SVM applied in a compact city model. Finally, conclusion and future work are offered in section VII .

B. Compact city model

The compact city model focuses on population density, open space protection, activity concentration, public transportation intensification, city size and access conditions, targeting socio-economic welfare [3] [4]. One of the first compact city models was developed by Swiss architect and urban designer LeCorbusier. According to him, the aim of this area is to concentrate high-density urban living associated with high-rise residential buildings. Recently the compact city model is an applied technique by scientists and urban planners. Around the world, this model is being used in different cities such as Amsterdam, Hamburg, Copenhagen, and in Japanese cities such as Wakkanai, Sendai, Toyama, Sapporo, Aomori, Toyohashi, Kobe, Kitakyushu and Fukuoka.

Japanese compact city model's aim could be defined in five main goals, namely:

1) Special attention at issue of aging,
2) Analysis and progress of suburbanization,

L.. Manrique is with the Computer Science and Engineering Department, University of Electro-Communications, Japan (e-mail: carlos@is.si.uec.ac.jp).

K. Yamamoto is with the Computer Science and Engineering Department, University of Electro-Communications,Japan (e-mail:k-yamamoto@is.uec.ac.jp).

3) Preservation of city history and culture,
4) Conservation of nature and environment,
5) Identification of the current status and future of regional collaboration.

The habitable area in Japan is less than 21% of its landmass and 66% is forest [5]. For that reason Japanese planners must think how to improve quality of life. For instance, in 2013 Aomori city had set aside about 985 million yen in its budget to clear snow for major roads, however more than 1.36 billion yen has been spent. This situation makes the government tries to bring together all the residents in the urban area, preventing urban sprawling, dealing with depopulation and aging (because life expectancy reaches almost 83 years old, whereas it is 10 years longer than in other developed countries), and thereby investing financial resources adequately. However, architectural problems present a risk on the citizens, because there are still wooden houses close to new buildings and it is important to improve the foundations in order to prevent disasters.

C. Related works

The SVM is a supervised non-parametric statistical learning technique, which has an important property: The determination of the model parameters corresponds to a convex optimization problem, where a local solution is also a global optimum.

Recently, SVM is applied in different kind of studies. For instance Zhou et al. [6] presented a method of Japanese dependency structure analysis based on SVM and conditional random fields (CRF). Their experiments demonstrated that combining SVM and CRF outperforms the cascaded chunking model based on sole SVMs and sole CRFs. Sudha and Bhavani [7] compared the efficiency between the k-Nearest Neighbor models (kNN) and multi class SVMs where multiple gait components were fused for enhancing classification rate. Their results demonstrated that the classification method using SVM was better than kNN. Pitiranggon et al. [8] developed a decompositional rule extraction technique from SVM called Support Vector Space Expansion (SVSE) rule. They applied it to financial data to predict currency crises. Ramirez-Gutierrez et al. [9] developed a face recognition algorithm using eigenphases and histogram equalization. They proposed a featured extraction scheme using SVM, the recognition rate was higher than 97% and verification error lower than 0.003%.

However, relevant studies related with land use are in remote-sensed data. Plaza et al. [10] focused on the methodologies for processing a specific type of imagery using SVM. Zhu and Blumberg [11] analyzed the different results of satellite image data applying different kernels such as radial basis and polynomials. Foody and Mathur [12] studied the potential for intelligent training sample collection; it was applied in classification of agricultural crops from multispectral satellite sensor data. They could classify crops with 92.5% of accuracy. Provost and Fawcett [13] discussed about the importance of the area under the Receiving Operating Characteristics (ROC) curve (AUC). They developed a hybrid classifier for any target conditions; the model is based on a method for comparison of classifier performance. This is done by combining techniques from ROC analysis, decision analysis and computational geometry. Brefeld and Scheffer [14] discussed an approximation for large data sets that clusters the constraints. Developing an AUC maximizing kernel machine they optimized a bound on the AUC and a margin term.

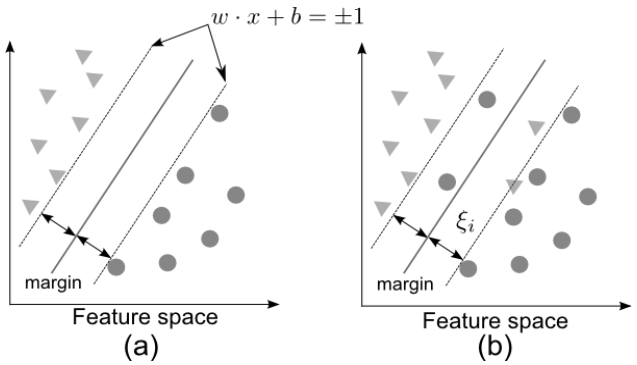

Figure 1. Optimal separating hyper plane. (a) separable samples, (b) non-separable data samples.

II. Method

A. Haversine distance calculation

By using GIS we calculated the haversine distance from each polygon to the different socio-economic factors. The area of each polygon is 1 ha and it is measured by the government each 100m. The haversine distance is defined as follows:

$$a = \sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos(\varphi_1) * \cos(\varphi_2)\sin^2\left(\frac{\Delta\lambda}{2}\right) \quad (1)$$

$$c = 2\tan 2\left(\sqrt{a}, \sqrt{1-a}\right) \quad (2)$$

$$d = R * c$$

Where: φ is latitude.
λ is longitude.
R is the earth's radius, defined here as (6,371 km).
d is the haversine distance.

B. SVM model

The SVM is considered as heuristic algorithms and it is based on statistical learning theory. The goal is to determine a hyper plane that optimally separates two classes.

Given a two separable classes with k samples defined as (x_i, y_i), where *i=1,2,....,k*, where $x \in R^n$ is an n-dimensional space, and $y_i \in \{+1, -1\}$ is a class label [15]. Suppose that the classes could be separated by two hyper planes parallel to the optimal hyper plane (Figure 1). The optimal hyper plane is represented as the line between the dotted lines.

$$w \cdot x_i + b \geq 1 \quad \text{for } y = 1 \text{ and } i = 1, \ldots, k. \quad (3)$$

$$w \cdot x_i + b \leq 1 \quad \text{for } y = -1 \quad (4)$$

Where $w=(w_1,....,w_n)$ is a vector of *n* elements.

Equations (3) and (4) can be combined into a single equation (5):

$$y_i[w' x_i + b] \geq 1, \text{ where } i = 1, \ldots, k. \quad (5)$$

The training data points on the hyper planes that are parallel to the optimal separated hyper plane (OSH) are the support vectors. The margin between the planes is defined as: $2/|w|$, the generalized linear SVM finds an OSH by solving the optimization problem :

$$\min\left[\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{r}\xi_i\right] \quad (6)$$

Subject to:

$$y_i[w\cdot x_i + b] + \xi_i - 1 \geq 0 \quad (7)$$

$$\xi_i \geq 0 \quad (8)$$

Where C is a penalty parameter on the training error, and ξ_i is the non-negative slack variable. The optimization model can be solved by introducing Lagrange multipliers for its dual optimization model [16]. If it is not possible to find a hyper plane by linear equations, the data must be mapped into a high dimensional space using nonlinear mapping functions (Φ).

$$f(x) = \text{sgn}\left(\sum_{i=1}^{k}\alpha_i y_i k(x, x_i) + b\right) \quad (9)$$

Where α_i is a Lagrange multiplier. $k(x, x_i)$ is a positive kernel that must meet Mercer's condition. Kernel functions can be aggregated into linear, polynomial, radial basis (Gaussian) functions and sigmoid kernels.

Linear kernel: $k(x_i, x) = (x_i, x)$ (10)

Polynomial kernel: $k(x_i, x) = (x_i \cdot x + 1)^d$ (11)

Where *d* is a natural number.

Radial basis function kernel:

$$k(x_i, x) = \exp\left(-\frac{1}{\sigma^2}\|x_i - x\|^2\right) \quad (12)$$

Sigmoid kernel: $k(x_i, x) = \tanh(kx_i \cdot x - \delta)$ (13)

By using the kernel function the nonlinear SVM classifier is defined as:

$$sign\left(\sum_{i=1}^{k}\alpha_i^* y_i k(x_i, x) + b^*\right) \quad (14)$$

SVMs were developed for binary classifications, however several studies such as one-against-all, one-against one and all together have been done for multiple class classification scenarios. In one-against-all a set of binary classifiers, each trained to separate one class from the rest [17]. One-against-one approach $k(k-1)/2$ SVMs are constructed for each pair of classes, the training data vector x_i is predicted to belong to the class with maximum number of votes.

C. Area Under the Curve (AUC)

The ROC allows assessing uncalibrated decision functions, even when the prior distribution of classes is unknown. The ROC curve details the rate of true positives against false positives over a threshold. The area of the ROC curve is the probability that a randomly drawn positive example has a higher decision function value than a negative example, it is called the AUC. The AUC is close to the Gini coefficient, the last one is used in random forest algorithms for classification variables.

The AUC is defined as follows:

$$AUC_{Total} = \frac{2}{|C|(|C|-1)}\sum_{\{c_i, c_j\}\in C}^{n} AUC(c_i, c_j) \quad (15)$$

Where *n* is the number of classes, $AUC(c_i, c_j)$ is the area under the two-class ROC curve between the classes c_i and c_j, and *C* is the set of all classes. The execution time of SVM [18] is calculated as:

$$T_{ex} = O(|C|^2 n \log n) \quad (16)$$

III. Outline of the Study Area

A. Study area description

Squires [19] defines MtA as a region with high population density in the core and a less populated perimeter, with shared industry, infrastructure and housing. The Statistics Bureau of Japan also defines MtA as one or more central cities which have social cohesion, special wards and ordinance designated cities and their surrounding municipalities. In Japan there are 14 MtAs, consisting of three major MtAs and other local MtAs. Local MtAs include Hokkaido, Tohoku, Hiroshima and Fukuoka regions.

In this study we select Aomori MtA as the study area, because it has been working as a compact city since 2000 and more than 90% of residents live in the urbanization promotion area (UPA). Aomori MtA is located between the geographical coordinates 139.5E to 141.2E and 40.4N to 41.1N in the northern hemisphere of Japan. This location is similar with some European countries such as Italy, Northern Greece, Albania, Portugal and France. Due to European geographical features it is possible to compare Japanese cities with European ones.

There are some areas detached from the main part of the MtA, these areas are defined as the commuter belt. One of the main characteristics of these areas is that all of them correspond to new town and they are near to a main road. In the

case of the southern part, there are three UPA recognizable outlying zones close to the Aomori prefectural route 44 which is part of the annular region. At northeastern part, there is one detached UPA which is close to the Aomori prefectural route 4 and Asamushionsen station.

B. Land use master plan of prefecture

In Japan, each prefecture has its own land use master plan consisting of a community facilities, traffic system, economic development, land use, parks and open space, neighborhoods and housing. There are different types of areas depending on the activity. Figure 2 represents the land use controls for city planning area and explains how city planning is defined by the local government. There are different city planning areas such as UPA, urbanization control area (UCA), district and zoning area. The UPA, according to the Ministry of Land, Infrastructure, Transport and Tourism (MLIT) of Japan, is defined as industrial, commercial and residential areas. While UCA is designated for agricultural activity and land use is regulated by plans. In 2000 there was a reform of zoning responsibilities. Nowadays the prefectures may freely decide whether to designate a zone or not. According to the percentage of areas with special land use controls for the Aomori MtA (Table 1), it is evident that the UPA (1.4%) is smaller than the UCA (11.2%), because it is focused on urban development projects. Figure 3 shows the UPA in the Aomori MtA [20].

C. Socio-economic factors

The studies related with socio-economic factors in Japan are focused on health, diet and mortality. Fukuda et al. [21] studied the sex-specific mortality of municipalities by age groups. They linked this problem with municipal socio-economic status (SES) indicators related to income, education, unemployment and living space. Their results showed that the mortality gradient had high impact on citizens less than 75 years old population than the total and over 75 years old, and the relationship between mortality and income-education related indicator was stronger for males than for females. The above mentioned authors continue studying the wide range of socio-economic factors associated with mortality focusing on factors such as unemployment, old housing, primary health resources and density. Their results showed that for women mortality, higher income, unemployment spacious dwelling, old housing, less vegetation, road facility number of cars per population, primary health resources and density were positively associated. Whereas higher education, public library activity and health check-up participation were independently negatively associated [22].

Other important socio-economic factors that cities have to deal with are: for instance, transportation and road systems, dwelling, industrial contamination of rivers, lakes, or coastal zones, degradation of landscape, shortage of green spaces and public recreation areas and lack of education, training, or effective institutional cooperation in environmental management [23]. According with the studies before mentioned, we will focus on transportation system such as bus stops and railroad stations, medical institutions, public facilities and we will study also land use price and other important facilities for the Japanese life style.

Table 1. Percentage of areas with special land use controls

City planning area	Percentage
UPA	1.4
UCA	11.2
Use district	0.9
Outside of zoning area	12.9
Inside city planning area	25.5
White area	13.3
Outside city planning area	74.5

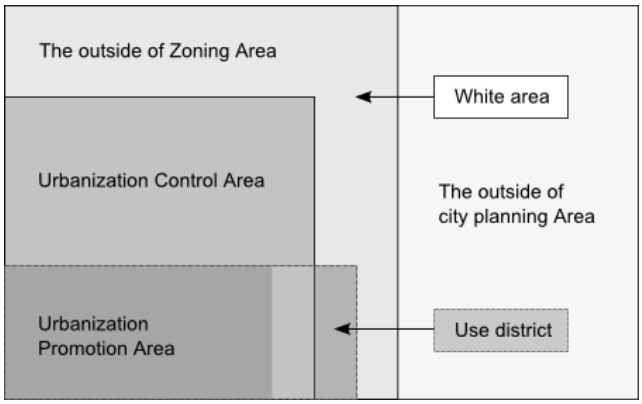

Figure 2. City planning area (Yamamoto [33])

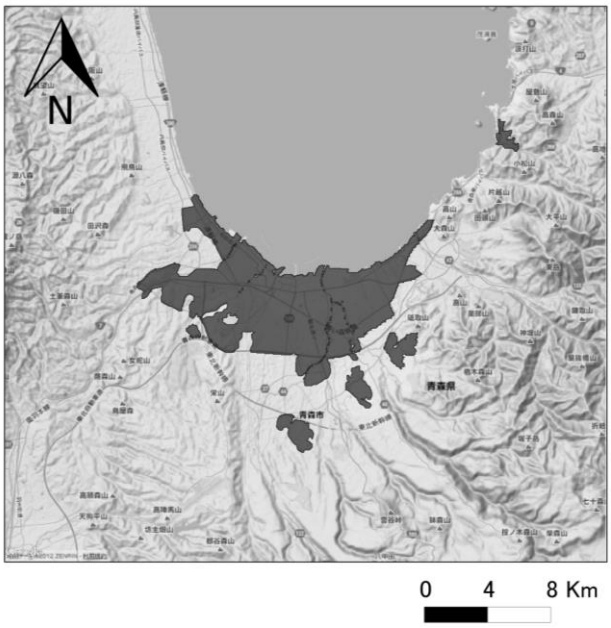

Figure 3. UPA in the Aomori MtA

IV. DATA PROCESSING

The information used for this study was, first of all, the related with the land use master plan of Aomori MtA was collected from the land use control back-up system (LUCKY) provided by MLIT. This system classifies different types of

Table 2. Land use categories

Code	Description	Color
1	Rice fields	
2	Other agricultural land	
3	Forest	
4	Waste Land	
5	Buildings, residential area	
6	Roads	
7	Other sites	
8	Rivers and lakes area	
9	Beach	
10	Ocean	
11	Golf course	

Table 3. Aomori's MtA (UPA)

Class	Area (ha)	Percentage(%)
Rice field	345	5.7
Other agricultural	140	2.3
Forest	206	3.4
Waste Land	45	0.7
Building site	3,848	63.8
Arterial traffic	305	5.1
Other	889	14.7
Rivers lakes	105	1.7
Beach	2	0.0
Ocean	146	2.4
Golf course	0	0.0
Total	6,031	100.0

Table 4. Socio-economic factors

Variable	Additional information	Total	Type
Bus stops	None	4,555	Num
Latitude, longitude	None	6,031	Num
Convenience stores	None	399	Num
Medical institutions	Hospital, Clinic, dental clinic, private schools, etc.	1,522	Num
Parks	None	754	Num
Land use price	By district	325	Num
Public Facilities	Building, national institutions, local government, schools, post office, etc.	3,216	Num Cat
Supermarkets	None	85	Num
Train stations	None	9	Num
Land use data	Urbanization promotion Area	6,031	Cat

Note: Numerical (Num), Categorical (Cat)

land use in the Japanese geography. In order to reduce calculation errors, it was necessary to analyze the image files (raster files) with an automated process [24] using GIS. These processes are regularly used in remote sensing, and allow the extraction of specific features. Through this system it is possible to extract the UPA.

Land use classification data was downloaded from the National Land Numerical Information download service also provided by MLIT for 2006. It is measured by 100 x 100m grids, and the classification system has a unique value per mesh area. Using GIS we overlaid the UPA's shape with the land use data to extract the land use information on the UPA. Table 2 shows the color system implemented in this study for each type of land use. However, in the UPA of Aomori MtA, there are not any golf courses. In Table 3 the information of the land use is shown, it is evident that building and residential area occupies more than 63% of the total area, while the area designated to other types of land occupies just 14.7%. According to UPA definition, the building and residential area should be promoted.

Socio-economic factors characterize the individual or group within the social structure. Among the most important socio-economic factors are education, income and occupation, place of residence, culture and ethnicity and religion. In this study, socio-economic factors that affect housing decision making were selected. These factors affect and define the activity in each grid area. The data related with the socio-economic factors such as railway stations, bus stops, convenience stores, malls, medical institutions, governmental and public services and land price by district was downloaded from the MLIT and public sources. Detailed information about them is shown in Table 4. The price of land is measured by district and it was needed to calculate it by 100 mt mesh areas. It can be defined as:

$$\Pr_a = \sum_{j=1}^{n} A_{a \in j} \Pr_j \qquad (16)$$

Where:

Pr_a is the price defined for the polygon a.

j are the districts intercepted by polygon a.

$A_{a \in j}$ is the area that shares a with the district j.

$\Pr_j$ is the price defined by the district j.

The *kernlab* package of the R programming language version 2.15 was used to calculate the SVM model. This tool is useful for kernel-based machine learning methods for classification, regression, clustering, novelty detection, quantile regression and dimensionality reduction. It also includes SVM, Spectral clustering, Kernel PCA. *Caret* package was also used to calculate the variable importance of the predictors, this package is also useful for data splitting, pre-processing, feature selection and model tuning using resampling.

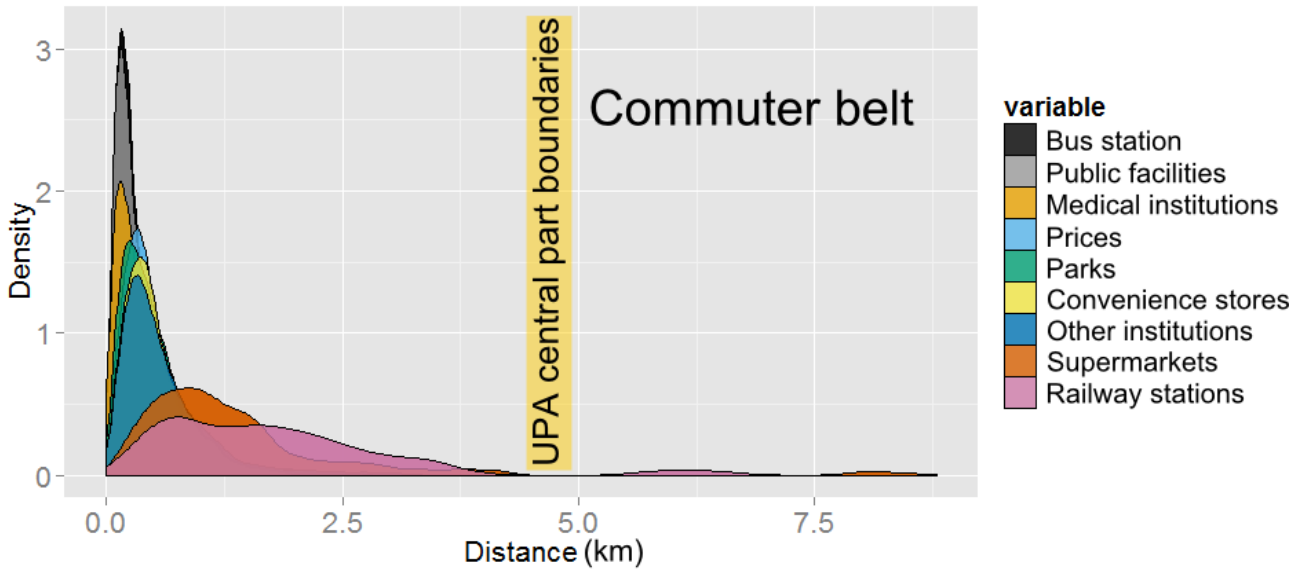

Figure 4. Distance from polygon to socio-economic factors (km)

Table 5. Experimental results

	Tuning 1	Tuning 2	Best model
Gamma value	0.01	0.1	0.07
Cost	760	7	100
Kappa Index	0.47	0.63	0.82
Rand Index	0.71	0.79	0.86
% diagonal values	0.74	0.81	0.90
Processing time (min)	19.90	93.75	>12 (hours)

V. RESULTS

A. *UPA analysis*

By calculating and rescaling the distance from each polygon's centroid to the socio-economic factors it was possible to evaluate the geographical distribution. Figure 4 shows density curve for distance from each polygon to socio-economic factors. According to the compact city definition, the boundaries have to be defined clearly, although the edges in the main part of the UPA are not defined particularly, it is possible to identify using geographical data that most of the social activities take place in this area. By calculating distances it was possible to identify the boundaries of the UPA central part at 5 km from the core, further than this distance there are few railway stations, supermarkets and other facilities. The database was randomly divided into two sets, 70% of the data was used as training and the remaining information was used for testing. The training set is used to train a multiclass SVM classifier [25].

B. *Parameterization*

According to literature on land use prediction models using SVM [26][27], the radial basis function Kernel presents a good performance in land use prediction. We have configured the SVM using this Kernel.

We have performed 3 different experiments to find a minimum training and cross errors, those experiments are called tuning 1, tuning 2 and best model. The parameters related with cost and gamma values are determined by grid search method using cross validation approach. The grid search method is useful for the computation of expensive numerical simulations and it has been applied in different studies in order to find a global minimum [28].

The first experiment consisted of 30 simulations, and total time was 19.90 min. The Kappa and Rand indexes were calculated to evaluate the SVM, the results were (0.47, 0.71) respectively. The Kappa index with larger values indicates better reliability, Kappa values greater than 0.7 are considered satisfactory. Rand index measures the percentage of decisions that are correct; it means that the prior results are still far from an accurate value. For that reason the experiments have to be improved. The number of vectors for this experiment was 2,386.

The computational time for the second experiment was 93.30 min, in this experiment we have run 200 simulations, the Kappa and Rand indexes were (0.63, 0.79) respectively, and the percentage of diagonal values was 81%, however the results are still far from good accuracy levels. The total number of vectors was 2,537. For those reasons we have extended the grid search in order to reach more accuracy and best model parameters.

The final experiment took more than 39 hours using a Windows based computer with 12Gb RAM memory, and processor Intel Xeon 2.67 Ghz. The number of simulations was 5015. In this experiment we found a Kappa and Rand Indexes (0.82, 0.86) respectively, the percentage of diagonal values is 90%. Results of the experiments are shown in Table 5. The standard deviation was calculated between the training error and cross error to choose the best simulation parameters. In this

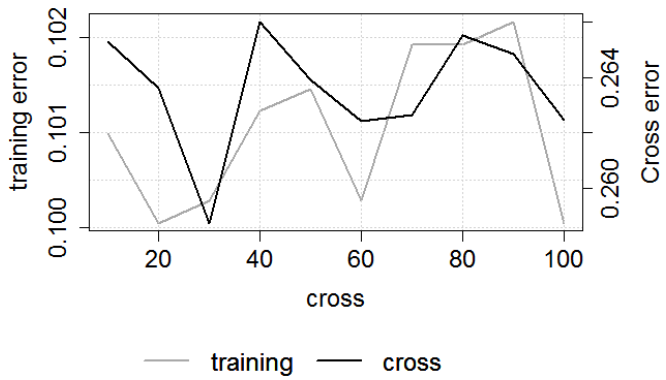

Figure 5. Results with cost same as 100

Table 6. Simulations' sample values

Cost	K-fold cross	Gamma	nsv	Tr.Er	Cr.Er	St.Dev
100	30	0.068	2429	0.100	0.259	0.112
100	70	0.067	2427	0.102	0.263	0.114
90	100	0.072	2431	0.101	0.262	0.114
100	60	0.068	2429	0.100	0.262	0.115
100	100	0.068	2429	0.100	0.262	0.115
100	50	0.067	2429	0.101	0.264	0.115
100	20	0.069	2430	0.100	0.264	0.116
100	80	0.067	2428	0.102	0.266	0.116
100	10	0.068	2428	0.101	0.265	0.116
100	40	0.068	2427	0.101	0.266	0.117

Note: **nsv**: Number of support vectors, **Tr.Er**: Training error, **Cr.Er**: Cross error, **St.Dev.**: Standard deviation.

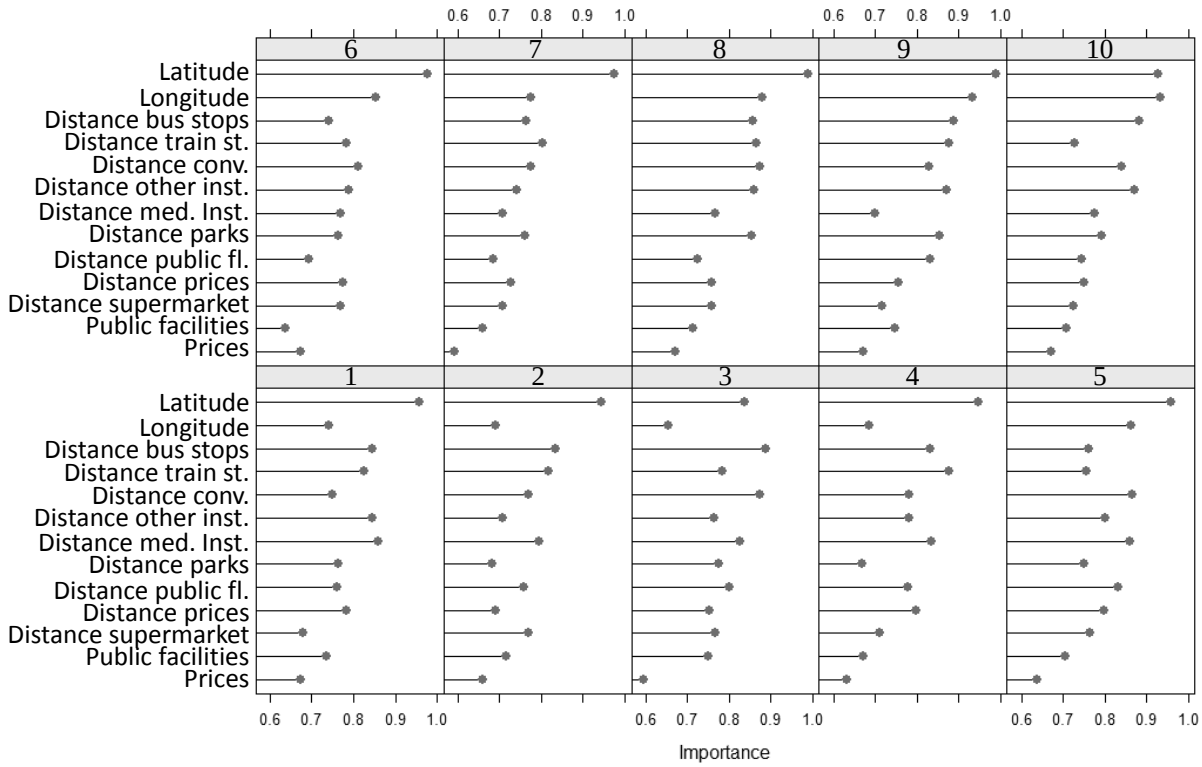

Figure 6. AUC for each predictor.

Note: **st**: Station, **med.**: medical, **inst.**: Institution, **fl.**: Facility, **conv.**: Convenience store

experiment, the cost value same as 100 presents the minimum training error (10.0%) when the cross value is same as 30 (Figure 5), Table 6 presents a simulation sample values. The number of vectors calculated was 2,424 with the same number of predictors. Although the first experiment presents the minimum number of vectors, the third experiment shows the best Kappa and Rand indexes.

C. AUC calculation

The third experiment was used to calculate the AUC for each predictor (Figure 6). The results for building and residential area parameter show that latitude is the factor that contributes most to the model with an area higher than (0.95). The factors given by the distance to other institutions, public facilities, medical institutions, longitude and convenience stores conformed the second group with an AUC value higher than (0.8). Finally, in a third group there are factors related with transportation methods, such as distance to train stations and bus stops, also the public facilities categories, parks, supermarkets. For all the factors the "latitude" has the largest AUC with a value higher than 0.9. And price of land is the factor that aggregates less information to the model.

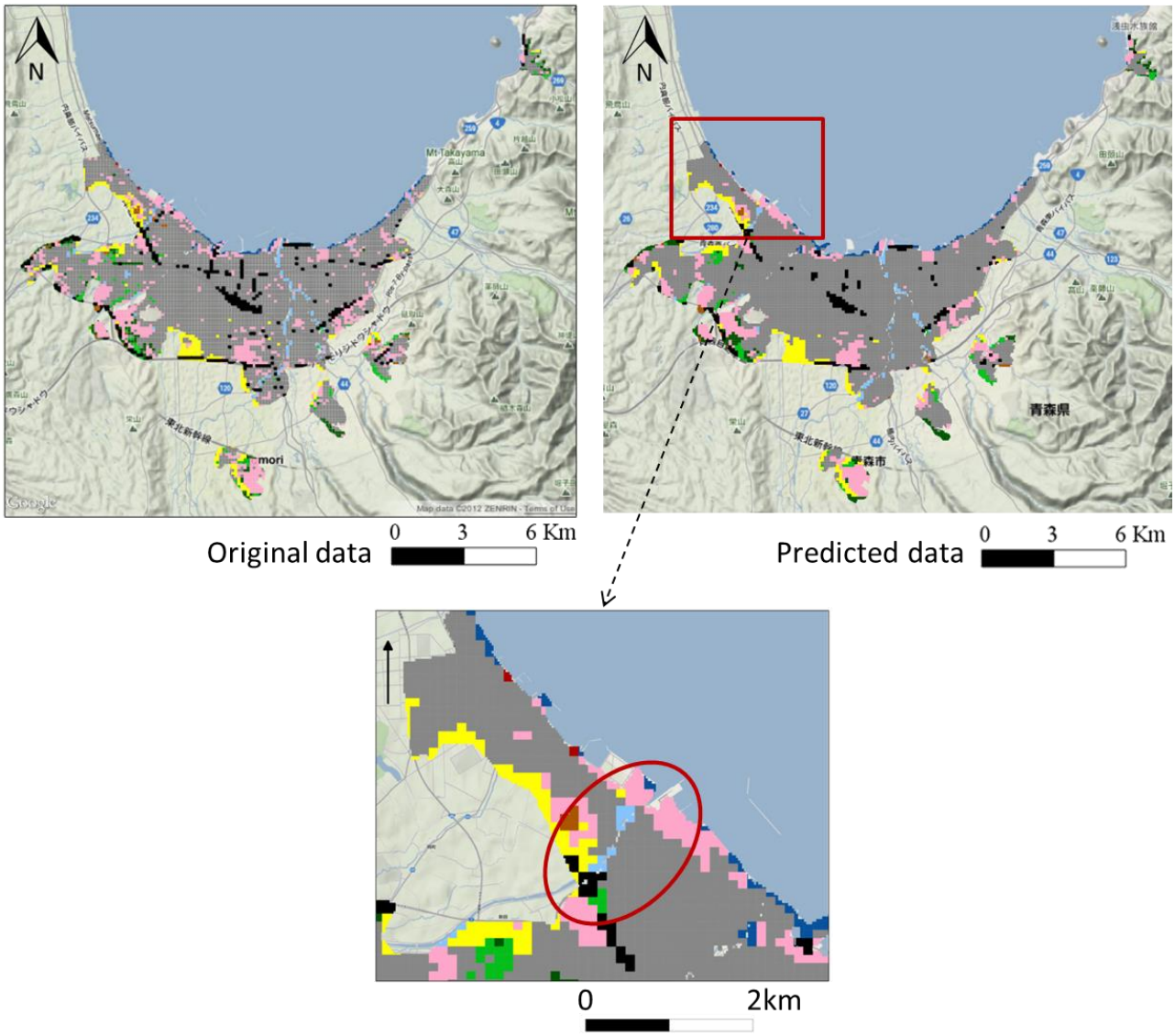

Figure 7. Aomori's UPA (detailed area)
Note: The top-left map illustrates the original information. Bottom map shows a detailed area across rivers surrounded by a circle. The largest errors on the SVM were on this class.

Figure 7 shows original and predicted information of UPA for Aomori MtA, also a detailed area across rivers is shown surrounded by a circle. The largest errors were on waste land (21.4%), across rivers and lakes (25.6%). However, the error close to beach area was (0.0%), forest (6.4%), buildings and residential area (6.7%).

VI. Discussion

The land use prediction using socio-economic factors has a promising future due to the importance of the model. The MtA of Aomori corresponds to the UPA, for that reason it was possible to identify the boundaries of this area. According to the compact city characteristics, the boundary has to be clear in order to bring people together and optimize resources such as energetic and transport. By using GIS we identify that more than 90% of people live in the UPA.

Rescaling data was needed in order to avoid errors in the classification process. That process allowed identifying the boundaries of UPA central area, because the density curves of the socio-economic factors do not present information between 4 km and 5 km. One of the most important characteristics of Aomori MtA is the commuter belt. This area affects the model's performance, because most of the socio-economic factors such as railway stations, shopping malls, hospitals and schools are located in the core of the MtA. Model calculation shows that this situation produces variability from the commuter belt to the core area.

Although the SVM is a useful technique for classification, the machine time consumption was expensive; it could be seen in the different experiments. However, in the third experiment appropriate Kappa and Rand indexes could be found. The experiment took 10 times more than the second one, however

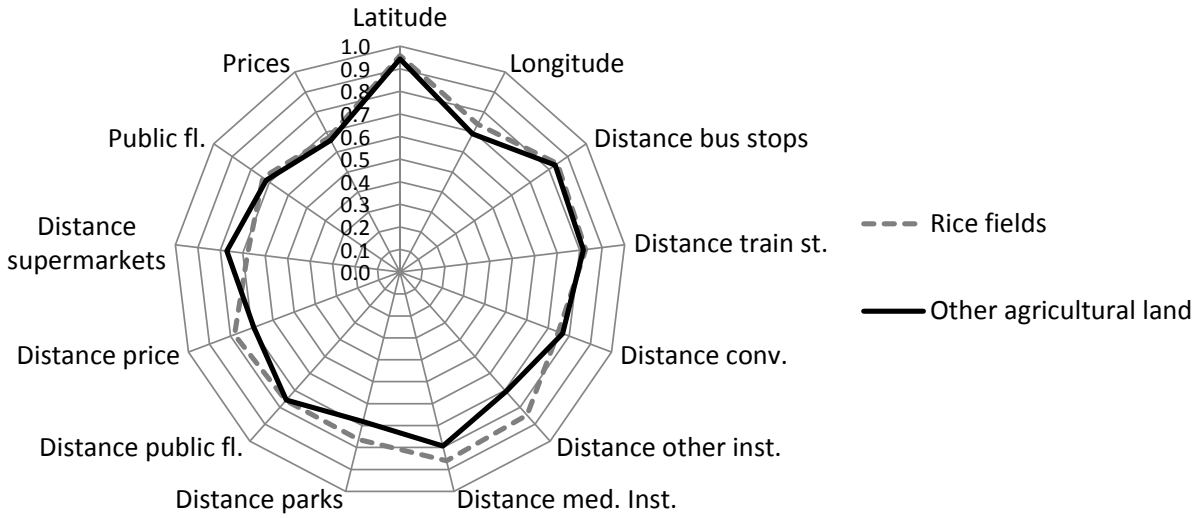

Figure 8. AUC comparison for rice fields and other agricultural land
Note: **st**: Station, **med.**: medical, **inst.**: Institution, **fl.**: Facility, **conv.**: Convenience store

the minimum global was found when the cost was 100 and gamma value same as 0.07. The grid search method was useful to identify the best values. In our experiments we verify that, when the cost was 100 it was possible to identify accurate values for the model, this result agrees with Liu et al. [29] by optimizing parameters of SVM model. In their study the gamma value was 0.143, with a Kappa value same as 0.84, for 5 land use classes. However in our study, the overall accuracy was 84% using 11 land use classes. The accuracy for roads and rivers classes was (66% and 74%) while beach, rice fields, forest and residential areas was higher than 93% and can reach 100%.

Simulations were important to define the model in the beginning, although the time machine required was considerable. Through grid search it was possible to identify parameters for each combination, however a local minimum was found without finishing all the experiments. Time computing for simulations was expensive; the experiments show that by increasing the computational time it was possible to find a minimum global. This is one of the simulation issues; however it is possible to find appropriate parameters once fitting values are acquired. In these experiments the conditions were in accordance with Kappa and Rand indexes, as well as training and cross errors.

AUC values were calculated for each predictor. The residential area shows that latitude presents an AUC over 0.96, distance to other institutions, public facilities, medical institutions, convenience stores and polygon's longitude have an AUC higher than 0.8 . According to the problems related with aging and depopulation, it is evident that health institutions and public facilities are more important than other factors. In a third group it is possible to identify that the AUC values for the groups related with transport, such as distance to railway stations and bus stops, distance to parks, distance to the closest district given prices and type of public facility are higher than 0.71. Finally the socio-economic factor that is less important is referent to prices with an AUC value same as 0.63.

AUC results for rivers and ocean predictors were similar; the standard deviations between the predictors for each factor was lower than 0.09, it means that the behavior between the two predictors is similar. The differences between the AUC values were compared and the P-value was higher than 0.1. It means that it is not significantly different. Rivers and ocean predictors depend on the geographical conditions, for that reason it is possible to get rid of both of them, and rather analyze the beach area taking into account the land reclamation.

Comparison between rice fields and other agricultural land (Figure 8) shows that the largest deviation was in distance to other institutions, distance to supermarkets and distance to the closest district to assign a land price. The comparison between the AUC values showed that the P-value is higher than 0.5; it reflects that the difference between both AUC values is not significantly different. This shows that rice fields and agricultural land parameters are similar, and both can be combined into one parameter to reduce the SVM model.

However, forest parameter should not be joined with the last group, because the P-value between the difference of forest, rice field and other agricultural land is lower than 0.05, and the human activity in the forest area is different than the other mentioned classes. If the classes such as rice field and agricultural land are joined, and rivers and lakes classes are joined in another group, the parameters could be reduced to 8 classes and the model accuracy could be improved.

New approaches have been done in order to optimize the computational time. The parallel computation for SVM [30] presents an approach to improve the number of calculations in order to reach the best parameters. The number of cross validations could be reduced at least one order of magnitude more than other grid search methods. By using parallel processing 60% more of function evaluations can be evaluated.

However, in order to use appropriately the parallel computing it is needed to think about the number of processors –also called workers- without job. Khun [31] showed that using the parallel processing the computational time using 10 processors is more than 5 times faster than in a serial computing; also the speedup is more than 10 times faster, where speedup is defined as the time for serial execution divided by the parallel execution time. It suggests that the parallel processing is necessary to calculate appropriate parameters. However, the maximum possible speedup achieved by parallelization with P number of processors is equal to P. For this study it may be necessary to reduce the system to 8 classes in order to gain efficiency by sacrificing speedup.

The prediction of land use cover and land use has been done using satellite image data by classifying the color information [32]. The data related with the different classes of land use are embedded into each vector file, for that reason the problem related with the color threshold and its correct classification does not exist in our study.

VII. Conclusion

In this study we have contributed to integrate geospatial information with socio-economic factors in order to classify land use. The SVM algorithm is useful to classify land use using information of socio-economic factors in a compact city model. By tuning the model, we could find the best cost and gamma parameters. It is important to choose the best variables which affect the housing in order to understand clearly the internal situation of the city. The UPA serves its purpose by gathering most of the residents in this area. It is possible to see the implementation of the compact city model in the UPA of Aomori MtA through the land use master plan and UPA's boundaries definition.

The use of the AUC helped to understand how each classifier affects the model. Using this criterion it is possible to understand hidden characteristics about the housing decision-making process. It was clear the influence of commuter belt through the AUC calculation. The most important variables for residents are related with transportation, health, land price and services.

In further experiments we will take special attention to the residents who live in the commuter belt. We will analyze what would happen in the case that the boundary of the Aomori compact city model was completely close (without detached areas). An approach of the SVM to middle and large scale of MtA will be given. In order to compare the compact city model behavior and its development, we will study other MtA which share similar characteristics. We will study their behavior taking 3 different periods of time, and in this way it is possible to understand the development of the core area and commuter belt.

Model parameters might be improved to increase the accuracy by extending the grid search or using an optimization method, also it is convenient to revise the mathematical formulation in order to reduce the time machine consuming. New variables will be added to improve the classification model and we will examine new relationships and ratios between them.

References

[1] L. Hermes, D. Frieauff, J. Puzicha and J. Buhmann, "Support vector machines for land usage classification in landsat TM imagery," in *Proc. Geoscience and Remote Sensing Symposium, 1999. IGARSS' 99*, Hamburg, 1999, pp.348–350.

[2] S. Berling-wolff and J. Wu, "Modeling urban landscape dynamics: A review," *Ecological Research*, vol. 19, pp. 119–129, 2004.

[3] M. Roychansyah, K. Ishizaka and T. Omi, "Considerations of Regional Characteristics for Delivering City Compactness: Case of Studies of Cities in the Greater Tokyo Area and Tohoku Region, Japan," *Journal of Asian architecture Building engineering*, vol. 4, no.2, pp. 339–346, Nov. 2005.

[4] L. Thomas and W. Cousins, "The compact city: a successful, desirable and achievable urban form?" in *The compact city, a sustainable urban form?*, M. Jenks and E. Burton, Ed. Oxford: E&FN Spon, 2010, pp. 53–65.

[5] N. Kadomatsu, "Recent Development of Decentralization, Deregulation and Citizens' Participation in Japanese City Planning Law," Kobe University law review, vol. 40, pp. 1–14, 2006.

[6] H. Zhou, T. Yu and D. Huang, "Japanese Dependency Analysis Based on SVMs and CRFs," *International Journal of Mathematics and Computers in Simulation*, vol. 1, no. 3, pp. 233–237,March 2007.

[7] L.Sudha and R. Bhavani, "Performance comparison of SVM and kNN in automatic classification of human gait patterns," *International Journal of Computers*, vol. 6, no. 1, pp. 19-28, 2012.

[8] P. Pitiranggon, S. Banditvilai and N. Benjathepanun, "Detection of Currency Crises by a Novel Rule Extraction Method from Support Vector Machine," *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 4, no.3, pp. 141-149, 2010.

[9] K. Ramirez-Gutierrez, D. Cruz-Perez, J. Olivares-Mercado, M. Nakano-Miyatake and Hector Perez-Meana, "A face recognition algorithm using eigenphases and histogram equalization," International journal of computers, vol. 5, no. 1, pp. 34-41, 2011.

[10] A. Plaza, J. Benediktsson, J. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image processing," *Remote Sensing of Environment*, vol. 113, no. 1, pp. 110-122, Sept. 2009.

[11] G. Zhu and D. Blumberg, "Classification using ASTER data and SVM algorithms; The case study of Beer Sheva, Israel," *Remote Sensing of Environment*, vol. 80, no. 2, pp. 233-240, May 2002.

[12] G. Foody and A. Mathur, "Toward Intelligent Training of Supervised Image Classifications: Directing Training Data Acquisition for SVM Classification," *Remote Sensing of Environment*, vol. 93, no. 1–2, pp. 107-117, 2004.

[13] F. Provost and T. Fawcett, "Robust Classification for Imprecise Environments," *Machine Learning*, vol. 42, no. 3, pp. 203–231, March 2001.

[14] U. Brefeld and T. Scheffer, "AUC maximizing support vector learning," in *Proc. ICML 2005 workshop on ROC Analysis in Machine Learning*, Bonn, 2005.

[15] C. Huang, L. Davis and J. Townshend, "An assessment of support vector machines for land cover classification," *International Journal of Remote Sensing*, vol. 23, no. 4, pp. 725–749, 2002.

[16] T. Oommen, D. Misra, N, Twarakavi, A. Prakash, B. Sahoo, and S. Bandopadhyay, "An objective analysis of support vector machine based classification for remote sensing," *Mathematical Geosciences*, vol. 40, pp.409–424. March 2008.

[17] A. Mathur and G. Foody, "Multiclass and Binary SVM Classification: Implications for Training and Classification Users," IEEE geoscience and remote sensing letters, vol. 5, no. 2, pp. 241–245, April 2008.

[18] P. Honzik, P. Kucera, O. Hyncica, and V. Jirsik, "Novel method for evaluation of multi-class area under receiver operating characteristic," in *Proc. Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control, 2009. ICSCCW 2009. Fifth International Conference on* , Famagusta, 2009, pp.1–4.

[19] G. Squires, "Urban sprawl: Causes, consequences and policy responses," Urban Institute Press, Washington, 2002.

[20] L. Manrique and K. Yamamoto, "Support vector machine for land use through socio-economic factors applied to a compact city model," in *Proc. System Science and Simulation in Engineering, 2013. ICOSSSE'13. Twelfth International conference on,* Morioka, 2013, pp. 113–120.
[21] Y. Fukuda, K. Nakamura, and T. Takano, "Municipal socioeconomic status and mortality in Japan: sex and age differences, and trends in 1973–1998," *Social Science & Medicine*, vol. 59, no. 12, pp. 2435–2445, Apr. 2004.
[22] Y. Fukuda, K. Nakamura, and T. Takano, "Wide range of socioeconomic factors associated with mortality among cities in Japan," *Health Promotion International*, vol. 19, no. 2, pp. 177– 187, June 2004.
[23] J. McCarthy, O. Canziani, N. Leary, D. Dokken and K. White, "Climate Change 2001: Impacts, Adaptation, and Vulnerability," Ed. Cambridge university press, 2001, pp. 403–405. [Online]. Available: http://www.grida.no/publications/other/ipcc_tar/?src=/climate/ipcc_tar/wg2
[24] C. Ratti and P. Richens, "Raster analysis of urban form," *Environment and Planning B: Planning and Design*, vol. 31, no.2, 297–309, 2004.
[25] D. Anguita, A. Ghio, L. Oneto, X. Parra and J. Reyes-Ortiz, "Human Activity Recognition on Smartphones Using a Multiclass Hardware-Friendly Support Vector Machine," in *Ambient Assisted Living and Home Care*, vol. 7657, J. Bravo, R. Hervás, M. Rodríguez, Ed. Springer Berlin Heidelberg, 2012, pp. 216–223.
[26] D. Shi and X. Yang, "Support Vector Machines for Landscape Mapping from Remote Sensor Imagery," in *Proc. AutoCarto 2012*, Ohio, 2012, pp. 16–18.
[27] J. Knorn, A. Rabe, V. Radeloff, T. Kuemmerle, J. Kozak and P. Hostert, "Land cover mapping of large areas using chain classification of neighboring Landsat satellite images," *Remote Sensing of Environment*, vol. 113, no. 5, pp. 957–964, 2009.
[28] I. Muntean and L. Dansorean, "Searching Simulation Scenarios on the Grid with ELSIGExplorer," in *Proc. 13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, Timisoara, 2011, pp. 261–267.
[29] Y. Liu, B. Zhang, L. Huang and L. Wang, "A novel optimization parameters of support vector machines model for the land use / cover classification," *Journal of Food, Agriculture & Environment*, vol.10, no.2, pp. 1098–1104. Apr. 2012.
[30] T. Eitrich and B. Lang, "Efficient Optimization of Support Vector Machine Learning Parameters for Unbalanced Datasets," *Journal of Computational and Applied Mathematics*, vol. 196, no. 2, pp.425–436, Nov. 2006.
[31] M. Khun, "Building Predictive Models in R Using the caret Package," *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, Nov. 2008.
[32] S. Prasad, T. Satya and I. Murali, "Classification of multispectral satellite images using clustering with SVM classifier", *International Journal of Computer Applications*, vol. 35, no. 5, pp. 32–44, Dec. 2011.
[33] K. Yamamoto and M. Nakamura, "An examination of land use controls in the lake Biwa watershed from the perspective of environmental conservation and management," *Lakes and Reservoirs: Research and Management*, vol. 9, no.3, pp. 217–228, 2004.

Luis Carlos Manrique Ruiz born in Bogota, Colombia in 1980. He received his degree on industrial engineer at La Sabana University, Colombia in 2007, graduated with honors. In march 2011 received the MS degree on information sciences focused on data mining from the Engineering Department of Gunma University, Kiryu, Japan. Currently, he is enrolled at the University of Electro-Communications in Tokyo, Japan. He is studying his doctoral course on Information sciences focused on city planning and he is expecting to finish on September of 2014. His interests are in spatial data mining, text-mining and real time GIS.

Kayoko YAMAMOTO received the B.H. Degree and M.H. Degree in Geography from Ochanomizu University in 1992 and 1994 respectively, and Ph.D. Degree in Social Engineering from Tokyo Institute of Technology in 1999. She is currently an associate professor in the Graduate School of Information Systems, National University of Electro-Communications, Tokyo, Japan. Her research interests include city planning and regional planning, environmental science and GIS (Geographic Information Systems).