SCIENCE OF THE TOTAL ENVIRONMENT 407 (2008) 358–378

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/scitotenv

Examining spatially varying relationships between land use and water quality using geographically weighted regression I: Model design and evaluation

Jun Tu[a,b,*], *Zong-Guo Xia*[c]

[a]*Department of Geography and Anthropology, Kennesaw State University, 1000 Chastain Road, Kennesaw, GA 30144-5591, USA*
[b]*Interdisciplinary Program of Environmental Studies, Kennesaw State University, 1000 Chastain Road, Kennesaw, GA 30144-5591, USA*
[c]*University of Massachusetts-Dartmouth, 285 Old Westport Road, North Dartmouth, MA 02747-2300, USA*

ARTICLE DATA

Article history:
Received 26 March 2008
Received in revised form
22 August 2008
Accepted 9 September 2008
Available online 30 October 2008

Keywords:
Geographically weighted regression
Spatial non-stationarity
Spatial autocorrelation
Land use
Water quality

ABSTRACT

Traditional regression techniques such as ordinary least squares (OLS) can hide important local variations in the model parameters, and are not able to deal with spatial autocorrelations existing in the variables. A recently developed technique, geographically weighted regression (GWR), is used to examine the relationships between land use and water quality in eastern Massachusetts, USA. GWR models make great improvements of model performance over OLS models, which is proved by *F*-test and comparisons of model R^2 and corrected Akaike Information Criterion (AIC_c) from both GWR and OLS. GWR models also improve the reliabilities of the relationships by reducing spatial autocorrelations. The application of GWR models finds that the relationships between land use and water quality are not constant over space but show great spatial non-stationarity. GWR models are able to reveal the information previously ignored by OLS models on the local causes of water pollution, and so improve the model ability to explain local situation of water quality. The results of this study suggest that GWR technique has the potential to serve as a useful tool for environmental research and management at watershed, regional, national and even global scales.

1. Introduction

Land use changes happening in the process of agriculture, industrialization and urbanization can modify watershed cover characteristics that influence runoff quality and quantity. Numerous studies have been conducted worldwide to analyze the relationships between land use and water quality using statistical methods (Basnyat et al., 1999; Jarvie et al., 2002; Tong and Chen, 2002; Little et al., 2003; Woli et al., 2004; Mehaffey et al., 2005; Williams et al., 2005; Rodriguez et al., 2007; Xiao and Ji, 2007). The primary statistical method is ordinary least squares (OLS) regression. Water quality variables from multiple sampling sites are used as dependent variables, and the corresponding land use variables and some other landscape characteristics (e.g. soil) derived from the drainage areas of the sampling sites are independent variables in the regression.

Two basic assumptions of OLS are that model residuals are uncorrelated with each other (no autocorrelation) and have constant variance (homoscedasticity). Violations of the assumptions will reduce the efficiency of the regression and mislead the interpretation of model results (Hamilton, 1992). However, the assumptions of no autocorrelation and homoscedasticity are often not considered in water environment

* *Corresponding author*. Department of Geography and Anthropology, Kennesaw State University, 1000 Chastain Road, Kennesaw, GA 30144-5591, USA. Tel.: +1 678 797 2392; fax: +1 678 797 2443.
E-mail address: juntu2005@gmail.com (J. Tu).

0048-9697/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.scitotenv.2008.09.031

studies. Violations of the assumptions of OLS in spatial data will bring two problems: spatial autocorrelation and spatial non-stationarity.

Spatial autocorrelation is the situation in which the value of a variable at a location is related to the values of the same variable at the locations nearby (Zhang et al., 2004). Spatial autocorrelation might exist in water quality variables at multiple sampling sites. One sampling site may have more similar water quality to a site nearby than to another site far away because the nearer sites may be affected by similar human activities (e.g. pollution sources) and natural characteristics such as soil, geology, land cover, and climate. Especially for sampling sites within the same watershed or even along the same stream, water quality among those sites might be more similar than sites from different watersheds. If a study sets up sampling sites in this way, it will bring spatial autocorrelation to the regression. However, many previous studies did not consider this issue (Jarvie et al., 2002; Woli et al., 2004; Xiao and Ji, 2007). They had some sampling sites within the same watershed or even along the same stream, but still used the water quality data for these sites to do OLS regression. In addition, this way to set up sampling sites might also bring spatial autocorrelation to land use variables. Land use data for a sampling site are gotten from its drainage area which might contain the drainage areas for other sites along the same stream or within the same watershed, so land use variables for these sites are closely related to each other, which is spatial autocorrelation. However, most previous studies seldom considered and discussed this spatial autocorrelation issue existing in land used variables (Jarvie et al., 2002; Woli et al., 2004; Xiao and Ji, 2007).

Spatial non-stationarity means that the relationships between the independent and dependent variables are not constant over space (Fotheringham et al., 2002). In previous studies, the relationship between a land use type and water quality obtained by OLS was for the whole study area, and represented the average situation, assuming the relationship does not change over space. However, this assumption is not true when we compared the results from different studies. For example, concentrations of Na and Ca, showed significant positive correlations with percentages of urban and agricultural lands, while TN (Total Nitrogen) and TP (Total Phosphorus) had no significant correlations with these two land use types in a study in the Ipswich River watershed, Massachusetts, USA (Williams et al., 2005). In contrast, a study in watersheds of the State of Ohio, USA found that agricultural land had significant positive correlations with TN and TP but negative correlations with Na and Ca, whereas urban land including residential and commercial lands had significant positive correlations with these four water quality indicators (Tong and Chen, 2002). The above examples show that the relationships between land use and water quality vary significantly over space, because watershed characteristics and pollution sources are not the same in different regions.

In addition, OLS regression may hide some local relationships. For example, in a study area containing different types of watersheds including highly-urbanized watersheds and agriculture-dominated watersheds, various pollution sources for some water quality variables may bring mixed effect on the relationships, and sometimes no significant correlations can be identified (Tu et al., 2007). A global relationship for the whole study area obtained by OLS is not able to represent the varying relationships over space.

In recent years, a simple but powerful technique called geographically weighted regression (GWR) has been developed to explore the spatially varying relationships and to account for spatial autocorrelation (Fotheringham et al., 2002). This technique has been applied in some ecology (Shi et al., 2006), social (Farrow et al., 2005), and urban studies (Yu, 2006). GWR attempts to capture spatial variations by allowing regression model parameters to change over space. The local estimation of model parameters is obtained by weighting all neighboring observations using a distance decay function, assuming that the observations nearby have more influence on the regression point than the observations further away. Besides a global coefficients of determination (R^2) calculated for the GWR model, a set of local regression results including local parameter estimates, the values of t-test on the local parameter estimates, the local R^2 values, and the local residuals for each regression point (sampling site) are also generated. Therefore, GWR might serve as a useful tool to explore the spatially varying relationships between land use and water quality.

In this study we applied the GWR technique to examine the spatially varying relationships between land use and water quality in eastern Massachusetts, USA. To our knowledge, this study is the first reported to apply GWR in examining the impact of land use on water quality, and it is even the first to use this technique in water environment research.

The presentation of the results is divided into two parts. This paper (Part I) focuses on model design and evaluation. It attempts (1) to fit both OLS and GWR models to investigate the relationships between land use and water quality indicators, (2) to test and compare the improvement of GWR over OLS on model fit, and (3) to examine the advantages of GWR over OLS in exploring the spatially varying relationships. The emphasis in this paper is not so much to determine whether or not there are relationships between land use and water quality, but to examine if any interesting spatial variations in the relationships exist. Part II, presented in a companion paper (Tu and Xia, submitted to Science of the Total Environment), analyzes in more detail the spatially varying correlations of different water quality indicators with different land use indicators examined by GWR and reveals the different major pollution sources in different parts of the study area.

2. Study area

As shown in Fig. 1, the study area covers metropolitan Boston and its surrounding areas up to 80 km away from Boston, which consists of 15 watersheds in eastern Massachusetts as defined by the U.S. Geological Survey (USGS) and the Massachusetts Department of Environmental Management's (MADEM) Division of Water Resources including Nashua, Black Stone, Charles, Taunton, and Ipswich River watersheds (Simcox, 1992). The area is about 10,000 km^2 with a population of about 5.2 million. This area is more densely populated, urbanized and industrialized than most parts of New England. High percentage of land is developed including residential, commercial, and industrial lands in metropolitan Boston and

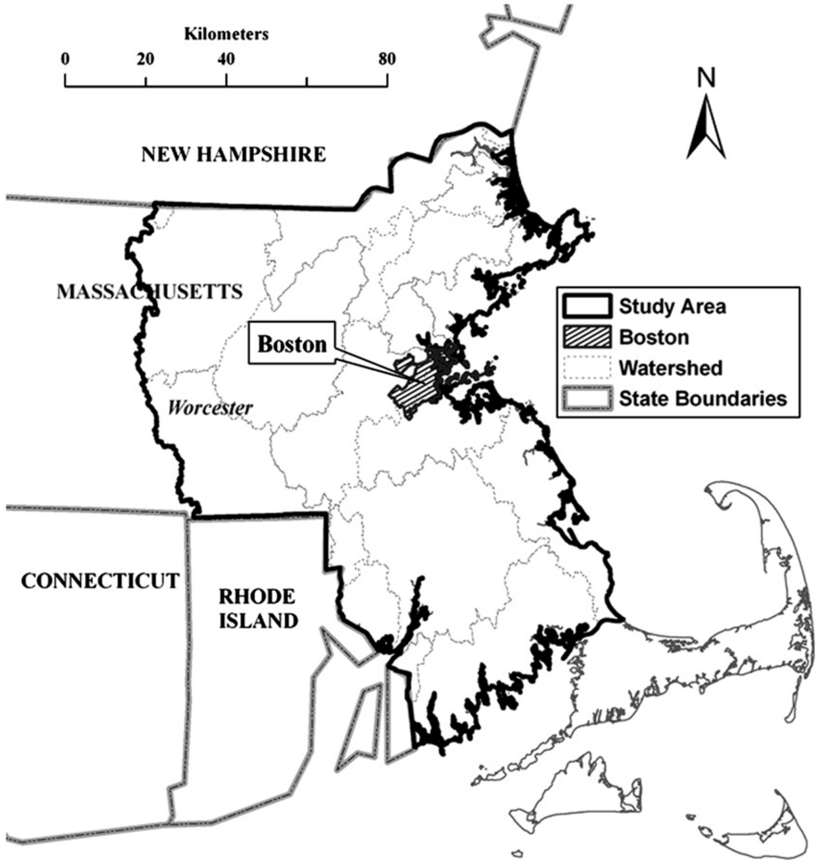

Fig. 1 – Location of study area.

Worcester, which can be classified as central cities and suburban areas. However, high spatial variability also exists. Developed land and population density decrease but forest and agricultural lands increase as the distance increases from the city of Boston. The population density in most of the municipalities within metropolitan Boston area is more than 4000 people/km^2, and the percentage of developed land is more than 80%. In contrast, the population density for many municipalities more than 20 km away from Boston is lower than 500 people/km^2, and the percentage of developed land is lower than 20%, while forest land may account for more than 80%, so these areas 20 km farther away from Boston can be considered as rural areas (Tu and Xia, 2006; Tu et al., 2007).

The land use change over time in this area is also considerable, and it also shows a great difference across space. Since the 1970s, the central city areas have experienced population decline and the land use change was relatively stable, whereas rural and suburban areas, especially the areas between 20 km and 40 km away from Boston, had substantial population growth and rapid land development (Tu et al., 2007). Developed land and population pushed their ways to all the directions except the ocean from the city of Boston, like a wave expanding from a center, which eroded forest and agricultural lands. Thus, land use changes caused by urban sprawl may bring great pressure on water resources. To evaluate the impact of land use changes on water quality and its potential spatial difference over the study area is very important for regional land use and watershed management.

It is well known that water quality in a watershed is also affected by many natural characteristics other than land use such as geology, physiography, vegetation, soil, and climate. The majority of the study area is within the U.S. Environmental Protection Agency (USEPA) Level III Ecoregion 59, the Northeastern Coastal Zone, which means that the watersheds are relatively homogeneous in their natural characteristics (Omernik, 1987). Therefore, performing the study in this area can minimize the impact of natural variability and allow us to focus on the spatial varying associations of water quality with land use.

3. Data sources and methods

3.1. *Water quality indicators*

Water quality data from 1990 to 2005 were retrieved on-line from the USGS National Water Information System Web (NWISWeb; URL http://waterdata.usgs.gov/nwis/). The NWISweb water quality data were collected by various projects ranging from national programs to projects in small watersheds. The methods of field sampling and measurements are documented in the USGS National Field Manual (USGS, variously dated). The laboratory analytical methods for the water quality parameters used in this study are described in the Techniques of Water-Resources Investigations of the U.S. Geological Survey (TWRIs) Book 5, Chapter A1 (Fishman and Friedman, 1989). The methods are consistent with the American Society for Testing and Materials (ASTM) standards and generally follow the standards of the International Organization for Standards (ISO). The quality assurance and quality control (QA/QC) procedures for laboratory analysis are described in TWRIs Book 5, Chapter A6 (Friedman and Erdmann, 1982). The Branch of Quality Systems in the USGS Office of Water Quality is in charge of monitoring, assuring, and improving the quality of analytical results. The accuracy of the water quality data provided to the public is insured by the USGS. The USGS is a very important and widely used public data source for research, education, and administration in the US.

The sampling sites, frequencies, methods, and water quality indicators are not designed for this study, so we had to rely on the existing data. Based on the data availability, 129 water quality sampling sites were selected, excluding the sites with part of upstream drainage area beyond the study area. Six dissolved ions, including calcium (Ca), magnesium (Mg), sodium (Na), potassium (K), chloride (Cl), and sulfate (SO_4), six dissolved nutrient parameters, including ammonia nitrogen (NH_3–N), nitrite nitrogen (NO_2–N), ammonia plus organic nitrogen (also known as kjeldahl nitrogen, KN), nitrate plus nitrite nitrogen (NO_3–N + NO_2–N), phosphorus (P), and orthophosphate phosphorus (PO_4–P), dissolved solid (residue on evaporation, DS), and specific conductance (SC), were used as water quality indicators.

The mean value of each water quality indicator over the whole study period at each sampling site was calculated. Not all the sampling sites had the data available for all the fourteen water quality indicators, so different sampling sites were used for different water quality indicators. For each water quality parameter, the sites with less than five samples were excluded from analyses. Thus, eventually, 94 sites for SC, 41 to 47 sites for dissolved nutrient parameters, and 31 to 37 sites for dissolved ions and solid were included for analyses (Table 1).

The varied sampling sites among different water quality indicators and the fact that data were collected in different projects would raise some doubts about the comparability of water quality among different sampling sites. However, we had to rely on the available historical data even though they are usually discrete and full of inconsistencies. The main reasons are described as follows. First, this study needs data from more sampling points, because a large number of sampling sites is very important for studying the spatial variability over a certain area by GWR. It is not feasible to set up many sampling sites with multiple water quality indicators over this study area during a short time period. Second, the purpose of this study is to test the advantages of GWR over OLS in exploring the spatially varying relationships between water quality and land use and to study the general pattern in the relationships. Therefore, as a pilot study of

Table 1 – Pearson correlations between water quality and land use indicators

Water quality indicator	n		AG	FO	OU	CO	IN	RE	RS	PDLU	PD
SC	94	r	***−0.366***	***−0.646***	0.154	***0.694***	***0.649***	***0.326***	***0.403***	***0.680***	***0.598***
		p	<0.001	<0.001	0.139	<0.001	<0.001	0.001	<0.001	<0.001	<0.001
NH_3–N	47	r	***−0.457***	***−0.688***	0.011	***0.536***	***0.396***	0.272	***0.421***	***0.665***	***0.717***
		p	0.001	<0.001	0.942	<0.001	0.006	0.064	0.003	<0.001	<0.001
NO_2–N	47	r	***−0.430***	***−0.688***	−0.171	***0.437***	0.182	***0.412***	***0.530***	***0.661***	***0.745***
		p	0.003	<0.001	0.250	0.002	0.221	0.004	<0.001	<0.001	<0.001
KN	42	r	***−0.401***	***−0.555***	−0.055	**0.377**	−0.010	**0.338**	***0.677***	***0.577***	***0.774***
		p	0.009	<0.001	0.730	0.014	0.950	0.029	<0.001	<0.001	<0.001
NO_3–N + NO_2–N	42	r	−0.098	**−0.394**	−0.057	0.296	0.211	0.246	0.279	***0.401***	***0.475***
		p	0.538	0.010	0.719	0.057	0.179	0.116	0.073	0.008	0.001
P	45	r	0.028	−0.149	**−0.329**	0.115	**−0.368**	0.214	**0.355**	0.153	***0.395***
		p	0.856	0.329	0.027	0.454	0.013	0.159	0.017	0.317	0.007
PO_4–P	41	r	0.304	0.293	0.074	−0.031	**−0.322**	0.012	0.007	−0.274	0.135
		p	0.053	0.063	0.644	0.847	0.040	0.938	0.965	0.083	0.400
Ca	32	r	**−0.406**	***−0.595***	0.021	***0.523***	***0.668***	−0.087	0.084	***0.656***	**0.393**
		p	0.021	<0.001	0.908	0.002	<0.001	0.635	0.646	<0.001	0.026
Mg	31	r	***−0.590***	***−0.679***	0.034	***0.573***	***0.592***	−0.073	0.256	***0.721***	***0.546***
		p	<0.001	<0.001	0.857	0.001	<0.001	0.697	0.165	<0.001	0.001
Na	32	r	***−0.557***	***−0.522***	0.179	***0.712***	***0.714***	−0.216	0.028	***0.589***	***0.458***
		p	0.001	0.002	0.327	<0.001	<0.001	0.235	0.878	<0.001	0.008
K	31	r	−0.212	**−0.410**	0.025	0.350	**0.417**	0.032	0.120	**0.429**	***0.571***
		p	0.253	0.022	0.892	0.054	0.020	0.862	0.522	0.016	0.001
Cl	37	r	**−0.405**	***−0.442***	0.146	***0.658***	***0.694***	−0.316	−0.082	***0.500***	**0.346**
		p	0.013	0.006	0.388	<0.001	<0.001	0.057	0.627	0.002	0.036
SO_4	32	r	−0.338	***−0.528***	−0.115	**0.442**	***0.492***	0.093	0.158	***0.586***	***0.473***
		p	0.059	0.002	0.529	0.011	0.004	0.613	0.386	<0.001	0.006
DS	34	r	***−0.543***	***−0.531***	−0.141	***0.550***	***0.464***	−0.023	0.206	***0.551***	***0.485***
		p	0.001	0.001	0.426	0.001	0.006	0.898	0.242	0.001	0.004

Note: AG = % agricultural land; FO = % forest land; OU = % open undeveloped land; CO = % commercial land; IN = % industrial, transportation, and mining land; RE = % recreation use; RS = % residential land; PDLU = % developed land; PD = population density. These abbreviations are also used for other tables.
Bold and italic number indicates values significant at $p < 0.01$.
Bold number indicates values significant at $p < 0.05$.

applying GWR in water environmental research, the quality and methods of processing the water quality data are acceptable.

The water quality data were also calculated by seasons and using median values and used for further analyses. However, the final results were basically the same as using the data over the whole study period and using the mean values. Thus, the results obtained by seasons and using median values are not reported in this paper.

3.2. *Land use indicators*

Land use for 1999 and population data by census block for 2000 were obtained from the website of the Massachusetts Geographic Information System (URL http://www.mass.gov/mgis/). The land use data were originally interpreted from 1:25,000 aerial photographs by the Resource Mapping Project at the University of Massachusetts, Amherst (MassGIS, 2005). Seven land use types were aggregated from 21 more detailed categories in the original data set by combining several similar land-use types into one broad category. The seven broad land use types were forest, agricultural land, open undeveloped land (defined as abandoned agricultural fields, power-line corridors, and areas of no vegetation), commercial land, industrial/transportation/mining areas (henceforth referred to as industrial land), recreation use, and residential land. Residential, commercial, industrial lands, and recreation areas were further aggregated into developed land use.

In order to get corresponding land use indicators for water quality sampling sites, the drainage area (sub-watershed) for each sampling site was delineated from digital elevation data provided by the USGS National Elevation Dataset (NED) 1 Arc Second (about 30 meter resolution, URL http://seamless.usgs.gov/website/Seamless/) using ArcGIS spatial analysis tools, and then used for the calculation of land use indicators.

Percentages of the seven land use types, percentage of developed land (PDLU), and population density (PD) were calculated in ArcGIS by overlapping land use and population layers to the drainage area of each sampling site and used as land use indicators for each water quality sampling site.

Through these steps, a linkage of water quality for points to land use indicators for areas was established. They are used for further statistical and spatial analyses.

3.3. *Modeling techniques*

According to the Kolmogorov–Smirnov test and visual interpretation of histogram and Q–Q plots, most of the land use and water quality indicators were not normally distributed. Thus, the variables that were not normally distributed were transformed using appropriate methods such as natural log and square root to meet the condition of normal distribution for further analyses.

Both OLS and GWR models were performed using water quality indicators as dependent variables and land use indicators as independent variables. In order to avoid the potential multicollinearity among land use variables, each of the OLS and GWR models used only one land use indicator as independent variable to analyze its association with each water quality indicator. There were nine land use indicators and fourteen water quality indicators in this study. Therefore, 126 OLS models and 126 GWR models were formed in this study.

Since OLS has been well known, we only briefly describe the theoretical background of GWR model in the following section. More detailed description of this technique can be found in some literatures (Brunsdon et al., 1998; Fotheringham et al., 2002).

3.3.1. *Theoretical background of GWR*

Geographically weighted regression is an extension of the traditional standard regression framework by allowing local rather than global parameters to be estimated (Fotheringham et al., 2001). It is a type of local statistics, which can produce a set of local parameter estimates showing how a relationship varies over space and then to examine the spatial pattern of the local estimates to get some understanding of hidden possible causes of this pattern (Fotheringham et al., 2002). In contrast, the traditional regression method such as OLS is a type of global statistics, which assumes the relationship under study is constant over space, so the parameter is estimated to be the same for all the study area.

An OLS model can be stated as:

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon \tag{1}$$

where y is the dependent variable, β_0 is the intercept, β_i is the parameter estimate (coefficient) for independent variable x_i, p is the number of independent variables, ε is the error term.

GWR model allows local rather than global parameters to be estimated for the location of sample and the above model can be rewritten as:

$$y_j = \beta_0(u_j, v_j) + \sum_{i=1}^{p} \beta_i(u_j, v_j) x_{ij} + \varepsilon_j \tag{2}$$

where u_j and v_j are the coordinates for each location j, β_0 (u_j, v_j) is the intercept for location j, β_i (u_j, v_j) is the local parameter estimate for independent variable x_i at location j.

GWR is calibrated by weighting all observations around a sample point using a distance decay function, assuming the observations closer to the location of the sample point have higher impact on the local parameter estimates for the location.

The weighting function can be stated using the exponential distance decay form:

$$w_{ij} = \exp\left(-d_{ij}^2/b^2\right) \tag{3}$$

where w_{ij} is the weight of observation j for observation i, d_{ij} is the distance between observation i and j, b is the kernel bandwidth. When the distance is greater than the kernel bandwidth, the weight rapidly approaches zero. Both fixed and adaptive kernel bandwidth can be chosen for GWR. Fixed kernel has a constant bandwidth over space, while adaptive kernel can adapt bandwidths in size to variations in data density so that bandwidths are larger in the locations where data are sparse and smaller where data are denser. We used adaptive kernel bandwidth in this study, because sample density varies over the study area. The optimal bandwidth was determined by minimizing the corrected Akaike Information Criterion (AIC_c) as described in Fortheringham et al. (2002).

GWR models produce a set of local regression results including local parameter estimates, the values of *t*-test on the local parameter estimates, the local R^2 values, and the local residuals, which can all be mapped to show the their spatial variability.

3.3.2. *Comparison between OLS and GWR models*

The purpose of comparing GWR with OLS models was to identify whether GWR models have better model performance than the corresponding OLS models. The comparison was performed by comparing the model R^2 and the AIC_c values from both GWR and OLS models.

Higher R^2 means that independent variable can explain more variance in dependent variable. A lower AIC_c value indicates a closer approximation of the model to reality, so lower AIC_c means better model performance (Wang et al., 2005).

In order to test whether the GWR models have a statistically significant improvement over the OLS models, an approximate likelihood ratio test, based on the *F*-test was performed (Fotheringham et al., 2002).

3.3.3. *Spatial autocorrelation calculation*

In order to compare the ability to deal with spatial autocorrelation between OLS and GWR models, global Moran's I was calculated for the residuals from each of the OLS and GWR models. Moran's I is a commonly used indicator of spatial autocorrelation. The value of Moran's I ranged from −1 to 1. A value of 1 means perfect positive spatial autocorrelation (high values or low values cluster together), a value of −1 suggests perfect negative spatial autocorrelation (a checkerboard pattern), and a value of 0 indicates perfect spatial randomness (Ishizawa and Stevens, 2007).

If significant spatial autocorrelation exists in an OLS model, then the model violates the assumption of randomly distributed and independent residuals in regression models. The efficiency of the model is therefore suspect. In addition, the residuals may contain some geographic information that the model does not include and therefore lost (Clark, 2007).

Global Moran's I statistics and Local Indicators of Spatial Association (LISA) analysis were also employed for each water quality indicator to explore the cause of spatial autocorrelation in OLS models. LISA measures the degree of spatial autocorrelation in each sample point by using local Moran's I. It is also good for identifying the existence of local spatial clusters by generating cluster maps (Longley and Tobón, 2004; Harries, 2006).

In this study, OLS models were performed using SPSS 13; GWR analyses were conducted using GWR 3 software package; Global Moran's I statistics and LISA were employed using GeoDa 0.9.5-i (Beta) analysis software (Anselin et al., 2006). All maps were made using ArcGIS 9.0.

Table 2 – Comparison of coefficient of determination (R^2) between OLS and GWR models

Water quality indicator		AG	FO	OU	CO	IN	RE	RS	PDLU	PD
SC	R_O^2	0.134	0.418	0.024	0.481	0.421	0.106	0.163	0.462	0.357
	R_G^2	0.632	0.769	0.597	0.759	0.841	0.740	0.406	0.559	0.792
NH_3–N	R_O^2	0.209	0.473	<0.001	0.288	0.156	0.074	0.177	0.443	0.514
	R_G^2	0.546	0.511	0.359	0.535	0.586	0.659	0.456	0.491	0.586
NO_2–N	R_O^2	0.185	0.473	0.029	0.191	0.033	0.170	0.281	0.436	0.554
	R_G^2	0.660	0.536	0.601	0.733	0.676	0.821	0.647	0.514	0.755
KN	R_O^2	0.160	0.308	0.003	0.142	<0.001	0.114	0.458	0.333	0.599
	R_G^2	0.805	0.902	0.618	0.848	0.850	0.845	0.870	0.905	0.806
NO_3–N + NO_2–N	R_O^2	0.010	0.155	0.003	0.088	0.045	0.061	0.078	0.161	0.226
	R_G^2	0.147	0.462	0.061	0.626	0.211	0.422	0.489	0.433	0.532
P	R_O^2	0.001	0.022	0.108	0.013	0.136	0.046	0.126	0.023	0.156
	R_G^2	0.033	0.162	0.510	0.670	0.531	0.672	0.714	0.169	0.629
PO_4–P	R_O^2	0.093	0.086	0.006	0.001	0.103	<0.001	<0.001	0.075	0.018
	R_G^2	0.660	0.348	0.566	0.652	0.621	0.664	0.652	0.355	0.582
Ca	R_O^2	0.165	0.354	<0.001	0.274	0.447	0.008	0.007	0.431	0.154
	R_G^2	0.462	0.793	0.500	0.759	0.731	0.685	0.763	0.818	0.725
Mg	R_O^2	0.348	0.461	0.001	0.329	0.350	0.005	0.066	0.519	0.298
	R_G^2	0.467	0.548	0.217	0.702	0.531	0.604	0.636	0.581	0.659
Na	R_O^2	0.309	0.272	0.032	0.506	0.510	0.047	0.001	0.346	0.209
	R_G^2	0.525	0.581	0.417	0.748	0.696	0.595	0.822	0.609	0.736
K	R_O^2	0.045	0.168	0.001	0.123	0.174	0.001	0.014	0.184	0.326
	R_G^2	0.189	0.864	0.065	0.598	0.263	0.477	0.515	0.808	0.515
Cl	R_O^2	0.164	0.195	0.021	0.432	0.482	0.100	0.007	0.250	0.120
	R_G^2	0.601	0.453	0.263	0.549	0.647	0.503	0.576	0.491	0.508
SO_4	R_O^2	0.114	0.279	0.013	0.195	0.242	0.009	0.025	0.343	0.224
	R_G^2	0.268	0.712	0.234	0.742	0.391	0.209	0.553	0.732	0.578
DS	R_O^2	0.293	0.279	0.020	0.305	0.212	<0.001	0.042	0.301	0.234
	R_G^2	0.390	0.674	0.215	0.469	0.319	0.685	0.855	0.701	0.629

Note: R_O^2 is the coefficient of determination for OLS model; R_G^2 is the coefficient of determination for GWR model.

4. Results and discussion

4.1. *Relationships between land use and water quality obtained by OLS*

The Pearson correlations between water quality and land use indicators obtained from OLS models are shown in Table 1. Except open undeveloped land, all the land use variables have significant correlations with most of the water quality indicators. Agricultural land and forest have significant negative correlations with SC, most of the dissolved nitrogen parameters, most of the dissolved ions, and DS, while commercial land has positive relationships with these water quality indicators. Industrial land exhibits significant positive correlations with SC, NH_3–N, all the dissolved ions, and DS, but significant negative correlations with the dissolved phosphorus parameters. Residential land and recreation use are significantly related to fewer water quality indicators. Residential land shows significant positive relationships with SC, NH_3–N, NO_2–N, KN, and P; recreation use has significant positive relationships with only SC, NO_2–N, and KN. The two composite land use indicators, PDLU and PD, have significant positive correlations with almost all the water quality indicators except the dissolved phosphorus parameters. Therefore, based on the results of OLS, PDLU and PD are the best predictors of water quality change over space.

However, as indicated by coefficient of determination (R^2) from OLS, neither PDLU nor PD can explain more than 60% of the variance in any water quality indicator (Table 2). The highest R^2 appears in the OLS model for PD and KN (R^2=0.599). Most of R^2 values are lower than 0.50, even for SC, which has

Table 3 – Statistical test results for improvement in model fit of GWR over OLS

Water quality indicator		AG	FO	OU	CO	IN	RE	RS	PDLU	PD
SC	AIC_O	**207**	**170**	**218**	**159**	**169**	**210**	**204**	**160**	**179**
	AIC_G	**145**	**160**	**165**	**115**	**120**	**145**	**179**	**133**	**147**
	p	**<0.001**	**<0.05**	**<0.001**	**<0.001**	**<0.005**	**<0.001**	**<0.001**	**<0.05**	**<0.005**
NH_3–N	AIC_O	184	165	195	179	187	192	186	168	161
	AIC_G	185	166	196	186	187	192	191	168	163
	p	>0.05	>0.05	>0.05	>0.05	<0.05	<0.05	>0.05	>0.05	>0.05
NO_2–N	AIC_O	484	463	**492**	**483**	**492**	**485**	478	466	455
	AIC_G	479	462	**485**	**472**	**479**	**463**	485	464	455
	p	<0.05	>0.05	**<0.05**	**<0.01**	**<0.01**	**<0.005**	>0.05	<0.05	>0.05
KN	AIC_O	**138**	**130**	**145**	**139**	**145**	**140**	**119**	**128**	107
	AIC_G	**128**	**117**	**129**	**129**	**126**	**119**	**116**	**117**	115
	p	**<0.01**	**<0.005**	**<0.005**	**<0.005**	**<0.001**	**<0.005**	**<0.05**	**<0.005**	>0.05
NO_3–N+NO_2–N	AIC_O	116	**109**	116	112	114	113	113	**109**	105
	AIC_G	114	**100**	116	113	114	115	113	**99**	103
	p	<0.05	<0.01	>0.05	<0.05	>0.05	>0.05	>0.05	**<0.005**	<0.05
P	AIC_O	224	223	219	223	217	**222**	**218**	223	**216**
	AIC_G	224	222	217	221	217	**218**	**209**	222	**207**
	p	>0.05	>0.05	>0.05	<0.05	>0.05	**<0.05**	**<0.05**	>0.05	**<0.05**
PO_4–P	AIC_O	428	429	**432**	**432**	**428**	**432**	**432**	**429**	**432**
	AIC_G	423	427	**423**	**425**	**425**	**427**	**422**	**426**	**422**
	p	>0.05	>0.05	**<0.05**	**<0.05**	**<0.05**	**<0.05**	**<0.05**	**<0.05**	**<0.05**
Ca	AIC_O	**104**	**96**	**110**	100	**91**	**110**	**110**	**92**	**105**
	AIC_G	**98**	**93**	**100**	99	**77**	**100**	**94**	**89**	**93**
	p	**<0.01**	**<0.05**	**<0.01**	<0.05	**<0.005**	**<0.01**	**<0.005**	**<0.05**	**<0.01**
Mg	AIC_O	34	28	**48**	35	**34**	47	45	25	**37**
	AIC_G	35	29	**44**	45	**31**	46	46	27	**34**
	p	>0.05	>0.05	**<0.05**	>0.05	**<0.05**	<0.05	<0.05	>0.05	**<0.05**
Na	AIC_O	**71**	73	**82**	60	**60**	**81**	83	69	75
	AIC_G	**66**	85	**74**	66	**53**	**78**	86	83	80
	p	**<0.05**	>0.05	**<0.01**	>0.05	**<0.01**	**<0.05**	<0.01	>0.05	<0.05
K	AIC_O	−8	−13	−7	−11	−13	−7	−7	−13	−19
	AIC_G	−7	3	−5	0	−12	0	1	−7	−19
	p	>0.05	<0.05	>0.05	>0.05	>0.05	>0.05	>0.05	<0.05	>0.05
Cl	AIC_O	**203**	202	**209**	119	**185**	**206**	209	199	205
	AIC_G	**190**	207	**204**	188	**181**	**201**	208	204	204
	p	**<0.005**	>0.05	**<0.01**	>0.05	**<0.05**	**<0.05**	<0.05	>0.05	<0.05
SO_4	AIC_O	44	37	47	**41**	39	47	47	34	**39**
	AIC_G	45	38	45	**35**	37	45	49	35	**34**
	p	>0.05	<0.05	<0.05	**<0.05**	>0.05	<0.05	>0.05	<0.05	**<0.05**
DS	AIC_O	−176	−176	−165	−177	−173	−164	**−165**	−177	−174
	AIC_G	−176	−154	−165	−179	−174	−159	**−172**	−157	−173
	p	>0.05	>0.05	>0.05	<0.05	>0.05	<0.05	**<0.005**	>0.05	>0.05

Note: AIC_O is the AICc for OLS model; AIC_G is the AICc for GWR model; *p* is the significant level for *F*-test; bold number indicates a significant improvement of GWR over OLS at $p<0.05$ and AIC_G is at least 3 less than AIC_O.

significant correlations with most of land use indicators. This result seems to be considerably different from the finding from a previous study in the same area (Tu et al., 2007). In that study, PD and PDLU could explain most of variance in dissolved ions and solid. For example, PD explained 92–97% of the spatial variation in DS, 85–94% in Cl, 73–98% in Ca, 73–91% in Na, 64–75% in Mg, and 45–79% in SO_4. The difference might be attributed to the different selections of studied sub-watersheds. In the previous study, the sub-watersheds were all mutually exclusive, which could avoid potential contamination by upstream sites on downstream sites. However, the selection strategy in that study limited the number of studied sub-watersheds. The number of sub-watersheds for SC in the previous study was only 18, and that for each of the other indicators was only around 10. Obviously, the limited number of studied sub-watersheds might reduce the significance of the findings from the study. In contrast, the number of sub-watersheds for SC in this study is 94, and the number for each dissolved nutrient parameter is more than 40, and that for each of the dissolved ions and solid is more than 30 (Table 1). Therefore, the significance of the findings from this study is expected to be improved greatly. Nevertheless, a lot of studied sub-watersheds in this study are mutually inclusive. In other words, one sub-watershed may contain another, and some water quality sampling sites come from the same stream. As we mentioned before, this selection strategy may bring spatial autocorrelation to statistical analyses and mislead the interpretation of the results of some traditional regressions (e.g. OLS). Thus, we are attempting to use GWR to reduce the spatial autocorrelation and also take advantage of high number of studied sub-watersheds. The results are shown and discussed in the following sections.

4.2. *Comparison of model performances between OLS and GWR*

The global R^2 of GWR with comparison of R^2 of OLS for each pair of dependent variable (water quality indicator) and independent variable (land use indicator) is given in Table 2. A dramatic improvement in R^2 of GWR over OLS is observed for every pair of water quality and land use indicators. The R^2 values for the models of PD and water quality indicators are improved from the values of 0.018 to 0.599 for OLS to the values of 0.508 to 0.806 for GWR. The R^2 values for PDLU in GWR range from 0.169 to 0.905, which is improved from the range of 0.023 to 0.519 in OLS. The ranges of the R^2s from GWR for commercial land, industrial land, recreation use, and residential land are 0.469–0.848, 0.211–0.850, 0.209–0.845, and 0.406–0.870, respectively. These values are all much higher than the corresponding R^2s from OLS. Likewise, agricultural land and forest also have much higher R^2 from GWR than that from OLS for each water quality indicator, even though they still have very low R^2 with phosphorus. Open undeveloped land, the weakest predictor in spatial variation of water quality found by OLS, also exhibits relatively high R^2 with almost all the water quality indicators (R^2=0.061–0.618 from

Table 4 – Comparison of Moran's *I* of the residuals from OLS and GWR models

Water quality indicator		AG	FO	OU	CO	IN	RE	RS	PDLU	PD
SC	I_O	**0.1057**	**0.1337**	**0.3649**	**0.1909**	**0.2450**	**0.2552**	**0.1555**	**0.0679**	**0.1461**
	I_G	−0.0180	−0.0310	−0.0166	−0.0436	**−0.0484**	−0.0337	−0.0102	−0.0303	−0.0372
NH3–N	I_O	**−0.1061**	−0.0123	−0.0124	−0.0570	−0.0645	−0.0311	−0.0343	−0.0352	−0.0598
	I_G	**−0.1147**	−0.0572	−0.0484	−0.0689	**−0.1069**	−0.0215	−0.0543	−0.0789	−0.0690
NO2–N	I_O	−0.0572	0.0472	0.0295	−0.0019	0.0082	−0.0026	−0.0047	0.0198	−0.0500
	I_G	**−0.1049**	−0.0233	−0.0076	−0.0538	−0.0721	−0.0005	−0.0708	−0.0468	−0.0343
KN	I_O	**−0.1081**	−0.0769	−0.0697	−0.0523	−0.0675	−0.0814	−0.0846	−0.0711	−0.0676
	I_G	**−0.1465**	**−0.1047**	**−0.1696**	−0.0399	−0.0638	**−0.1582**	−0.0337	−0.0638	−0.0581
NO3–N+NO2–N	I_O	−0.0441	**0.1067**	−0.0520	−0.0008	−0.0397	−0.0205	0.0256	**0.1041**	**0.0928**
	I_G	**−0.1005**	−0.0897	−0.0917	−0.0815	**−0.1311**	−0.0923	−0.0593	−0.0718	−0.0334
P	I_O	−0.0448	−0.0020	−0.0152	−0.0163	−0.0889	−0.0137	0.0351	0.0006	**0.0550**
	I_G	−0.0527	−0.0382	**−0.1223**	−0.0344	**−0.1035**	**−0.0858**	−0.0185	−0.0334	−0.0197
PO4–P	I_O	0.0415	−0.0412	**0.0806**	**0.0783**	0.0232	**0.0887**	**0.0879**	−0.0266	**0.1302**
	I_G	**−0.1128**	**−0.1500**	**−0.1183**	−0.0836	**−0.1375**	−0.0810	−0.0850	**−0.1388**	−0.0678
Ca	I_O	**0.2540**	**0.3108**	**0.3168**	**0.2497**	**0.2248**	**0.2254**	**0.2979**	**0.2254**	**0.3989**
	I_G	−0.0599	−0.0088	−0.1443	−0.0688	**−0.2106**	0.0029	−0.0496	−0.0477	−0.0571
Mg	I_O	−0.0409	−0.0517	0.0733	**0.1009**	−0.0150	0.0712	0.0513	−0.0808	**0.1896**
	I_G	**−0.1346**	−0.1015	−0.0670	−0.0688	**−0.1436**	−0.0728	−0.0652	−0.1236	−0.0700
Na	I_O	**0.1682**	**0.2041**	**0.2842**	**0.1810**	**0.1562**	**0.2791**	**0.2509**	**0.1298**	**0.2981**
	I_G	−0.0635	−0.0331	−0.0997	−0.0778	−0.1546	0.0008	−0.0124	−0.0562	−0.0583
K	I_O	**−0.1583**	−0.0482	**−0.1738**	−0.1247	**−0.1608**	**−0.1709**	**−0.1489**	−0.0570	0.0195
	I_G	**−0.2012**	−0.1013	**−0.2000**	−0.0874	**−0.2227**	−0.1146	−0.1101	**−0.1593**	−0.0402
Cl	I_O	**0.1965**	**0.1501**	**0.2131**	**0.1314**	**0.1140**	**0.2144**	**0.1999**	**0.1149**	**0.2277**
	I_G	−0.0486	−0.0416	0.0081	−0.0208	−0.0598	−0.0408	−0.0389	−0.0455	−0.0556
SO4	I_O	0.0439	**0.1258**	0.0682	0.0256	0.0169	0.0783	0.0634	0.0794	**0.1498**
	I_G	−0.0880	−0.0589	−0.1330	−0.0733	−0.1334	−0.0833	−0.0826	−0.0791	−0.0845
DS	I_O	−0.0443	−0.0543	−0.0294	0.0167	−0.0494	−0.0035	−0.0117	−0.0555	0.0362
	I_G	**−0.1178**	−0.0491	**−0.1340**	−0.0756	**−0.1323**	−0.0694	−0.0449	−0.0449	−0.0616

Note: I_O is the Moran's I for OLS model; I_G is the Moran's I for GWR model; bold number indicates values significant at $p<0.05$ level; bold and italic number indicates values significant at $p<0.01$ level.

GWR). Therefore, GWR models make considerable better predictions of water quality than the corresponding OLS models.

The higher values of the global R^2s from GWR than the R^2s from OLS indicate the improvement in model performance of GWR over OLS. However, the statistical significances of the improvements need to be checked by the values of AIC_c and F-test. The statistical test results for improvement in model fit of GWR over OLS are shown in Table 3.

Comparisons of the AIC_c values from multiple models with the same independent variable provide a relatively simple way to decide the best model. A lower AIC_c value indicates a closer approximation of the model to reality (Wang et al., 2005). Usually, a decrease of AIC_c lower than three might be caused by sampling error, but not a result of a genuine difference in models (Fotheringham et al., 2002). Thus, in this study, a GWR model is considered to be significantly improved from its corresponding OLS model if the AIC_c value of the GWR is at least three lower than that of the OLS and the F-test is significant at $p<0.05$ level.

As shown in Table 3, significant improvements in model fit of GWR over OLS are found for all models of SC, and most of the models for KN, PO_4–P, and Ca. Fifty-five of the total 126 GWR models have significant improvements over their corresponding OLS models. Among the remaining 71 GWR models, thirty-one have either significant F-test at $p<0.05$ level or lower AIC_c than OLS, so they are considered as non-significant improvements over OLS models. Only fifteen GWR models with AIC_c value of at least three higher than that for the corresponding OLS model and non-significant F-test. Therefore, even checked by AIC_c values and F-tests, most of the GWR models are still better than the corresponding OLS models in model fit.

4.3. *Comparison of spatial autocorrelations of residuals from OLS and GWR*

The results of Moran's I statistics on the residuals from OLS and GWR models are shown in Table 4. Significant positive spatial autocorrelations are found for all the OLS models for SC (Moran's $I=0.0679$–0.3649, $p<0.01$), for Ca (Moran's $I=0.2248$–0.3989, $p<0.01$), for Na (Moran's $I=0.1298$–0.2981, $p<0.05$ or $p<0.01$), and for Cl (Moran's $I=0.1140$–0.2277, $p<0.05$ or $p<0.01$). In contrast, almost all the corresponding GWR models for these four water quality indicators have no significant spatial autocorrelations in their residuals. This result indicates that these OLS models are unsuitable for identifying the relationships between land use and water quality indicators as the spatial dependences in water quality are not adequately accounted for by the OLS models and the OLS assumption of residual independence is not met, while GWR models improve the reliabilities of the relationships by reducing the spatial autocorrelations in residuals. Significant spatial autocorrelations are also found for some OLS models for some other water quality indicators including NH_3–N, KN, NO_3–N + NO_2–N, P, PO_4–P, Mg, K, SO_4, and DS. However, significant spatial autocorrelations still remain in the residuals for some GWR models for these water quality indicators, and even the spatial autocorrelations are increased from some OLS models to the corresponding GWR models. The latter case usually happens for the OLS models without significant spatial autocorrelation problem (e.g. some models for P and DS).

The differences in spatial autocorrelations of OLS model residuals might be explained by the different spatial distributions of water quality indicator values. The results of the global Moran's I statistics and LISA analysis are illustrated in Fig. 2. As shown in Fig. 2, significant positive spatial autocorrelations are found for SC (Moran's $I=0.1072$, $p=0.002$), Ca (Moran's $I=0.1707$, $p=0.002$), and Cl (Moran's $I=0.0933$, $p=0.03$). Correspondingly, significant positive spatial autocorrelations are exhibited in the residuals of OLS models for these three water quality indicators (Table 4). The cluster maps for these three water quality indicators, especially for SC, show several significant spatial clusters and clear spatial distribution patterns of the clusters. Most of high-high clusters, which indicate that high values of water quality indicators are surrounded by high values, are concentrated within the 20-km buffer of Boston, the highly-urbanized central city area. Low-low clusters, indicating that low values are surrounded by low values, are located in the south part of the study area outside the 20-km buffer, even outside the 60-km buffer, which are relatively less-urbanized rural areas. Low-high and high-Low clusters, the areas with the mix of low and high values of water quality indicators, are mainly within the areas between high-high and low-low clusters, which are mainly the interfaces between urban and rural areas (suburban areas). Na, which is another water quality indicator with significant spatial autocorrelations in the residuals for its OLS models, has slightly significant spatial autocorrelation (Moran's $I=0.049$, $p=0.07$), and its clusters have the similar spatial pattern.

On the contrary, the other water quality indicators, especially the dissolved nutrient parameters, have no significant spatial autocorrelations (Moran's $I=-0.0923$ to 0.0613, $p=0.06$ to 0.86) and fewer spatial clusters (Fig. 2). As a result, they have fewer significant spatial autocorrelations in the residuals of OLS models. If GWR is performed on some of these variables, the spatial autocorrelations may be increased.

Thus, the application of GWR requires a careful use of diagnostic tools such as Moran' I statistics. If an OLS model has spatial autocorrelation problem, GWR can help reduce it. On the other hand, if an OLS model has no this problem, an application of GWR may increase spatial autocorrelation. Nevertheless, for the purpose of comparing the overall improvement of GWR over OLS in this study, GWR still show its ability to deal with spatial autocorrelation. Fifty-six of 126 OLS models have significant spatial autocorrelations in their residuals, whereas only 28 GWR models have this problem (Table 4).

4.4. *Spatial varying relationships between water quality and land use explored by GWR*

Maps of local parameter estimates, t-test, and local R^2 from GWR models allow the visualization of the spatial varying relationships between land use and water quality indicators. We are not able to present all the maps here, as there are 126 GWR models in this study. Only the maps of the results from GWR models for SC are shown in this paper to describe the spatial non-stationarity of the impact of land use on water quality (Figs. 3–11). SC is a measure of the ability of water to conduct electrical current and reflects the concentrations of

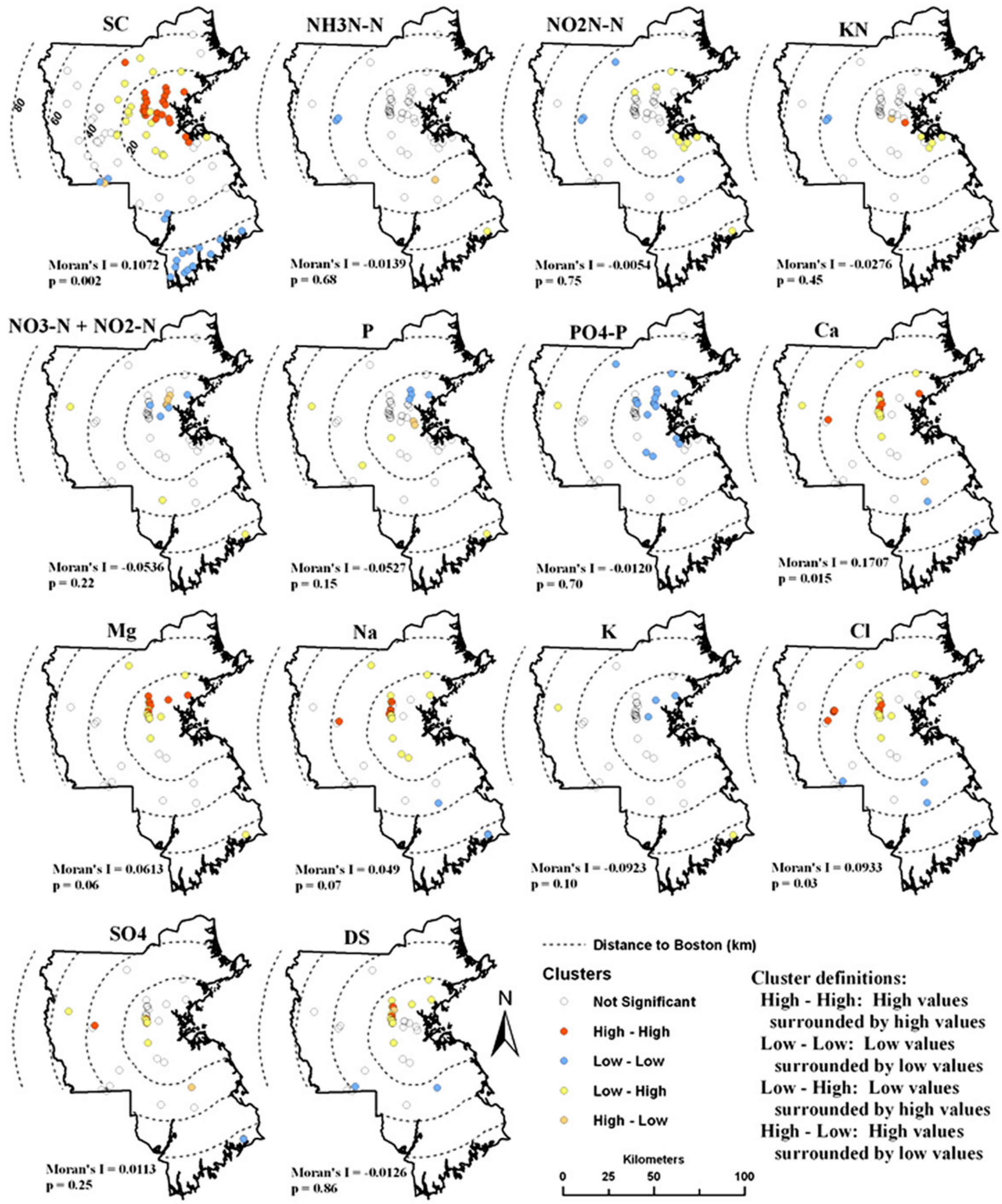

Fig. 2 – Cluster maps of water quality indicators.

dissolved ions or solid in water. SC is the water quality indicator most significantly affected by land use in this study identified by both OLS and GWR models (Tables 1 and 2). Because its strong correlation with human disturbance (e.g. land use), it was considered as one of the best general water quality indicators in many studies (Wang and Yin, 1997; Dow and Zampella, 2000; Tu et al., 2007).

Fig. 3 shows the maps of local parameter estimates, t-test, and local R^2 from the GWR model for SC and agricultural land. From the results of OLS model for SC and agricultural land we know that a significant negative correlation exists between SC and agricultural land (r=0.366, p<0.001; Table 1). However, Fig. 3a shows that both negative and positive correlations are distributed in the study area, and a clear spatial pattern can be identified. The areas within the 20-km buffer from Boston, mainly metropolitan area, and the southern and northern parts of the areas outside the 20-km buffer have negative correlations, indicating that higher percentage of agriculture land is related to lower concentration of SC. On the other hand, the western part of the area outside the 20-km buffer, which is mainly rural area except the city of Worcester, show positive correlations, suggesting that higher percentage of agriculture land is related to higher concentration of SC. In other words, in the western part of the area outside the 20-km buffer, agriculture land might make

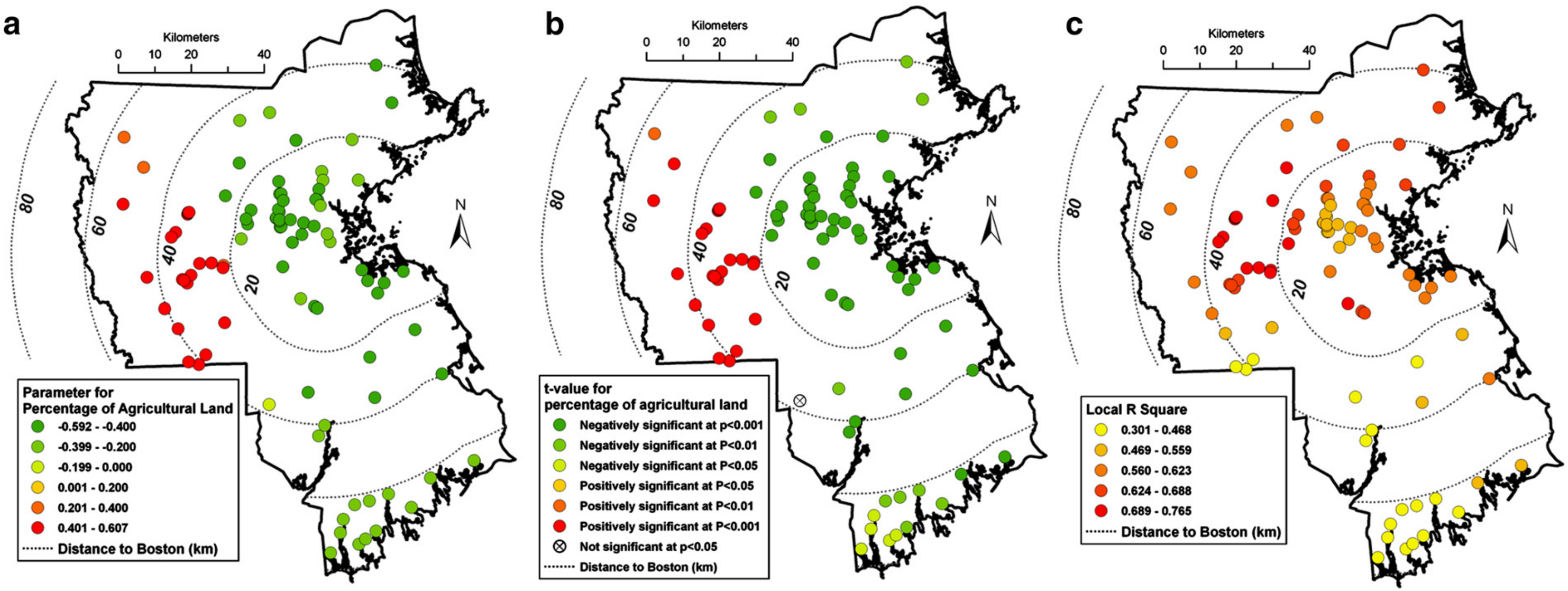

Fig. 3 – Results of the GWR model for SC and percentage of agricultural land: (a) parameter estimates for percentage of agricultural land; (b) *t*-value for percentage of agricultural land; (c) local R^2.

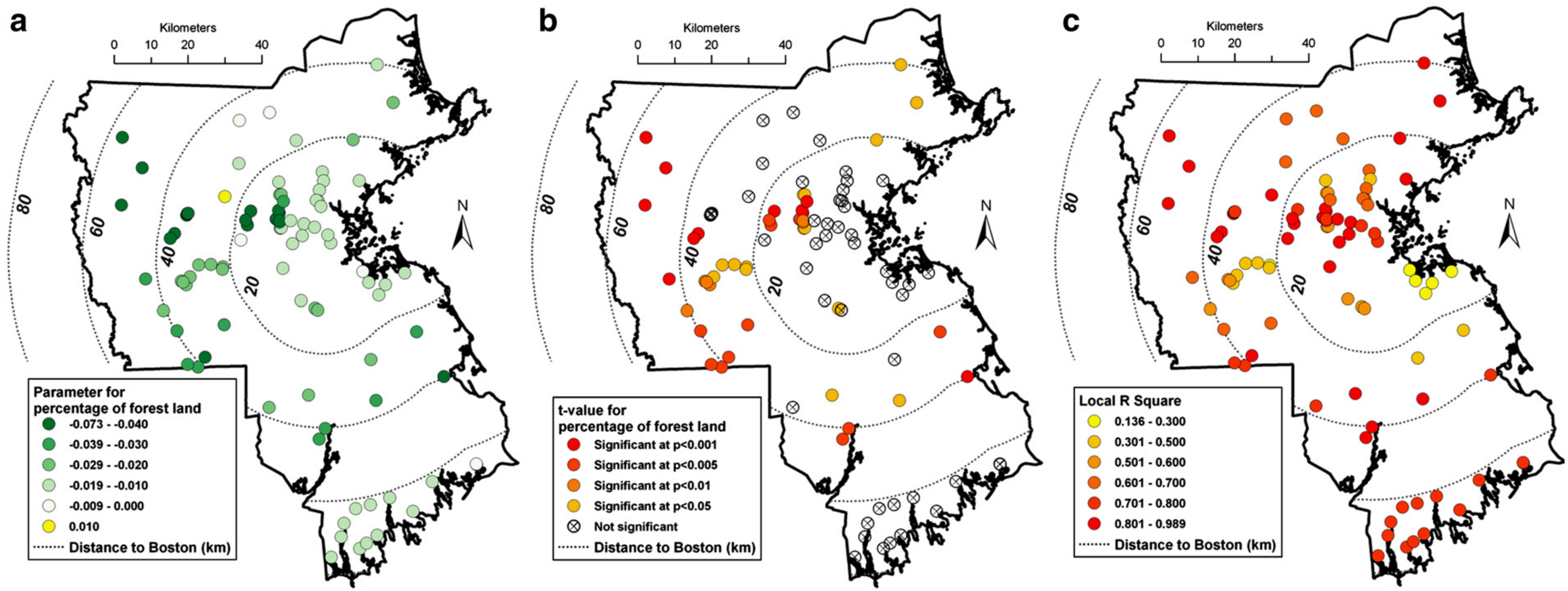

Fig. 4 – Results of the GWR model for SC and percentage of forest land: (a) parameter estimates for percentage of forest land; (b) *t*-value for percentage of forest land; (c) local R^2.

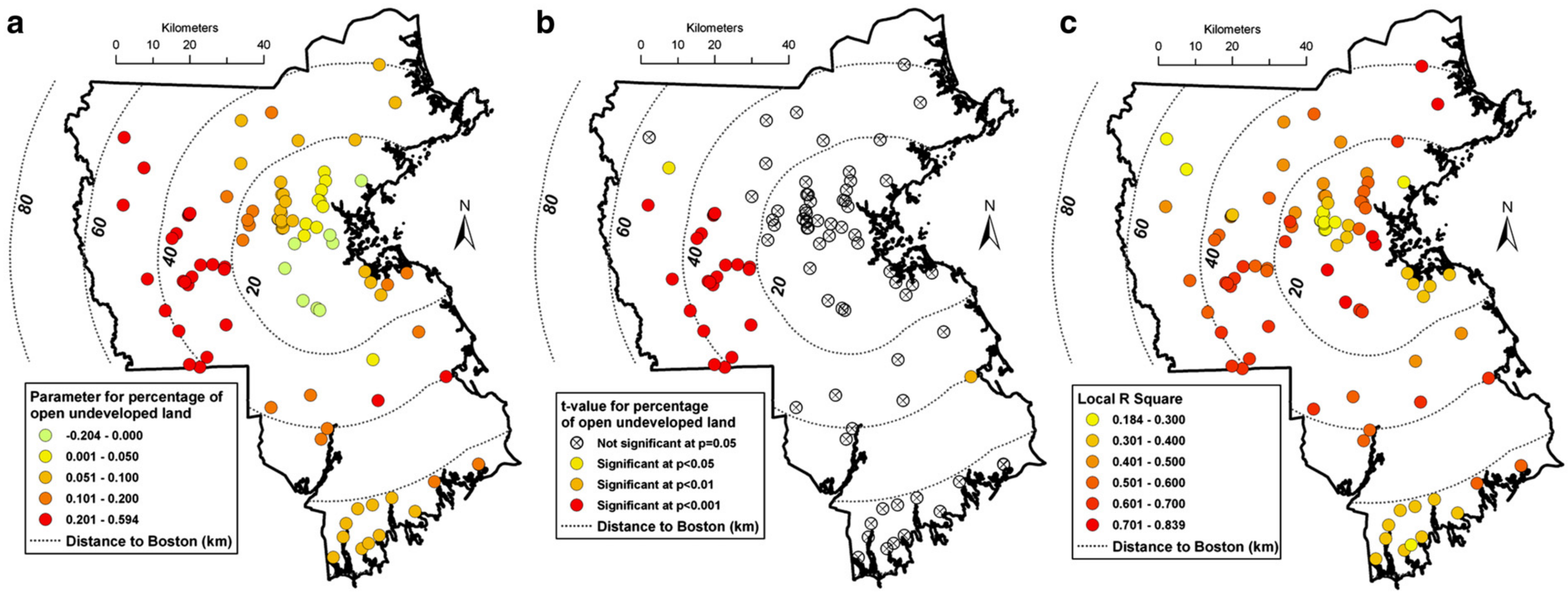

Fig. 5 – Results of the GWR model for SC and percentage of open undeveloped land: (a) parameter estimates for percentage of open undeveloped land; (b) *t*-value for percentage of open undeveloped land; (c) local R^2.

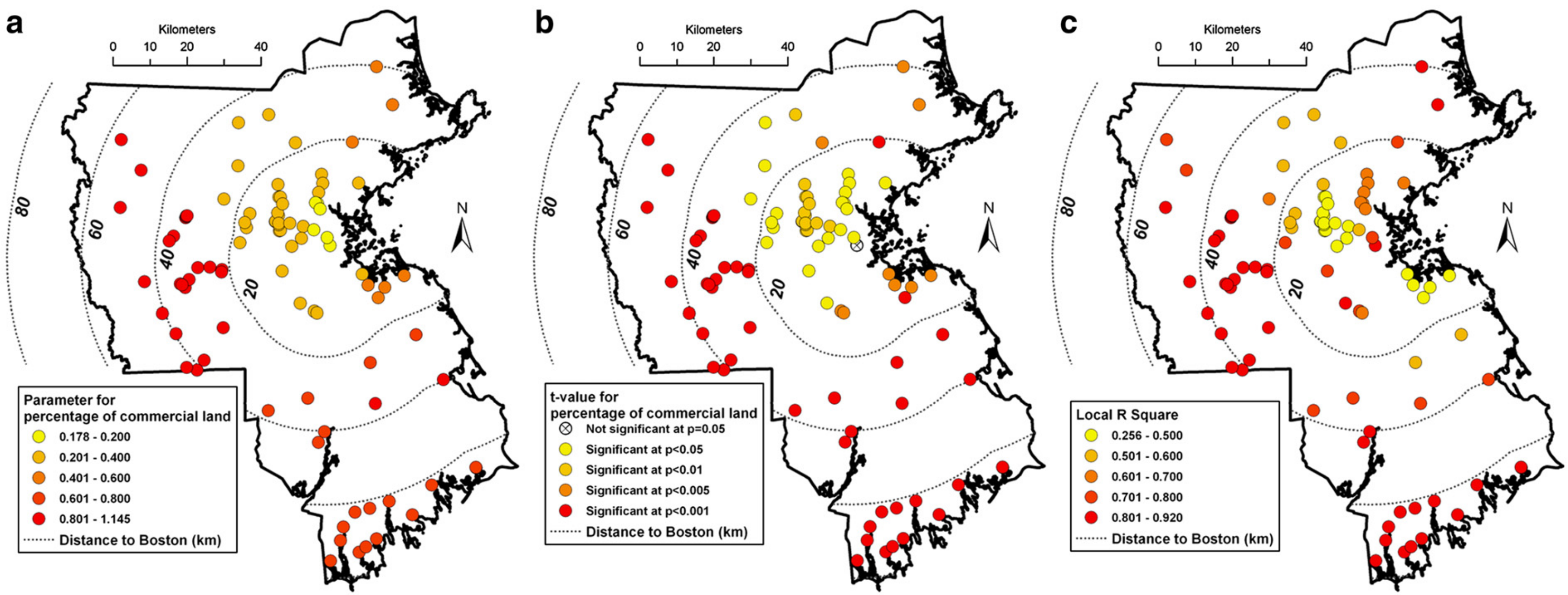

Fig. 6 – Results of the GWR model for SC and percentage of commercial land: (a) parameter estimates for percentage of commercial land; (b) *t*-value for percentage of commercial land; (c) local R^2.

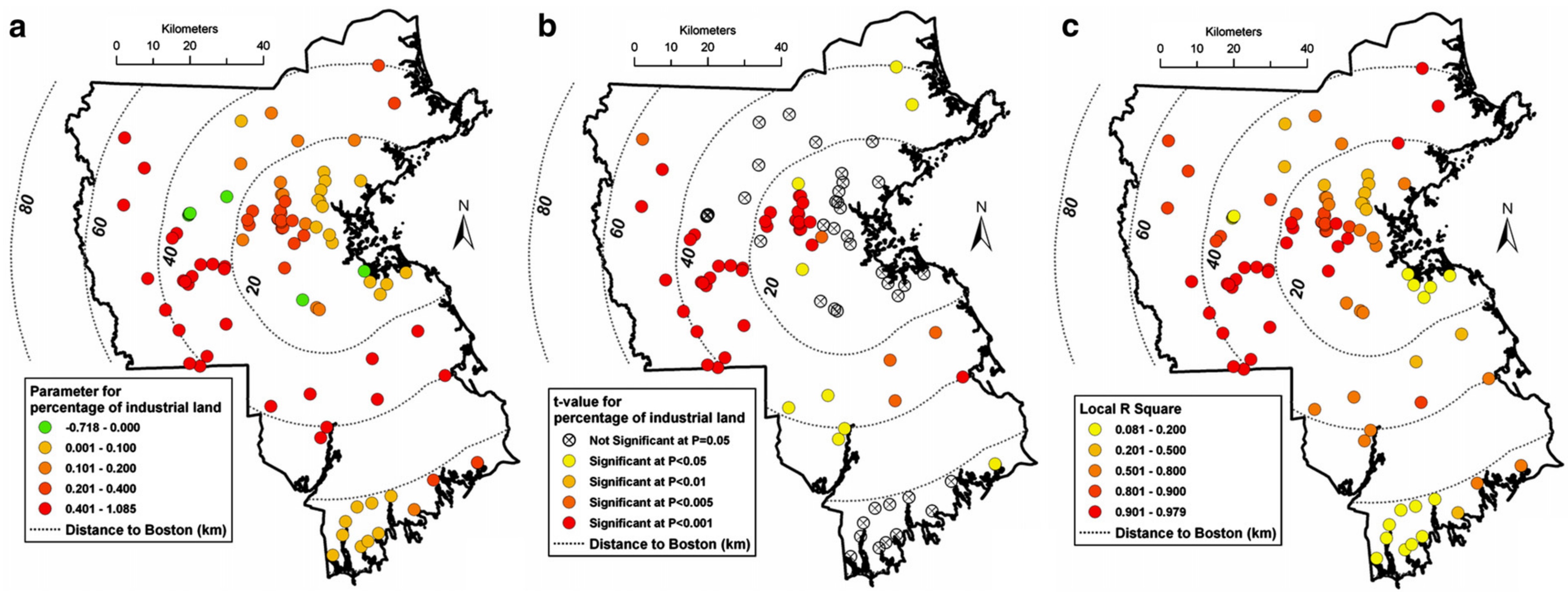

Fig. 7 – Results of the GWR model for SC and percentage of industrial land: (a) parameter estimates for percentage of industrial land; (b) *t*-value for percentage of industrial land; (c) local R^2.

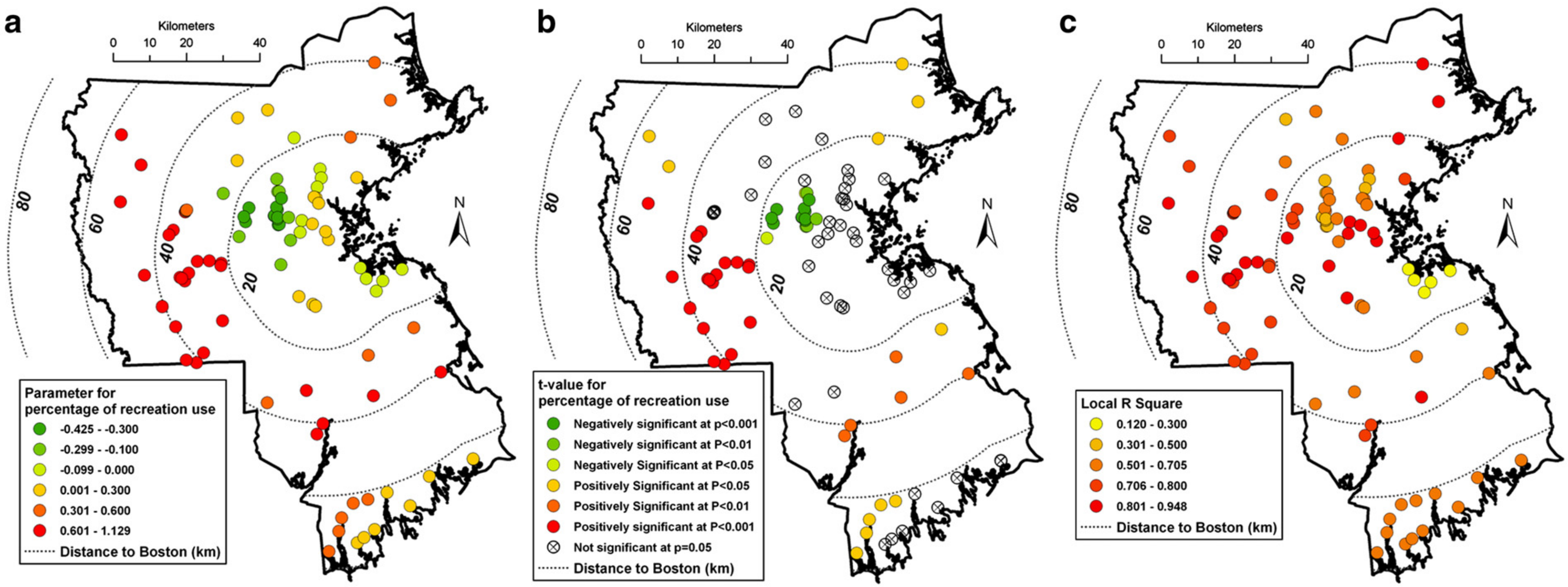

Fig. 8 – Results of the GWR model for SC and percentage of recreation use: (a) parameter estimates for percentage of recreation use; (b) *t*-value for percentage of recreation use; (c) local R^2.

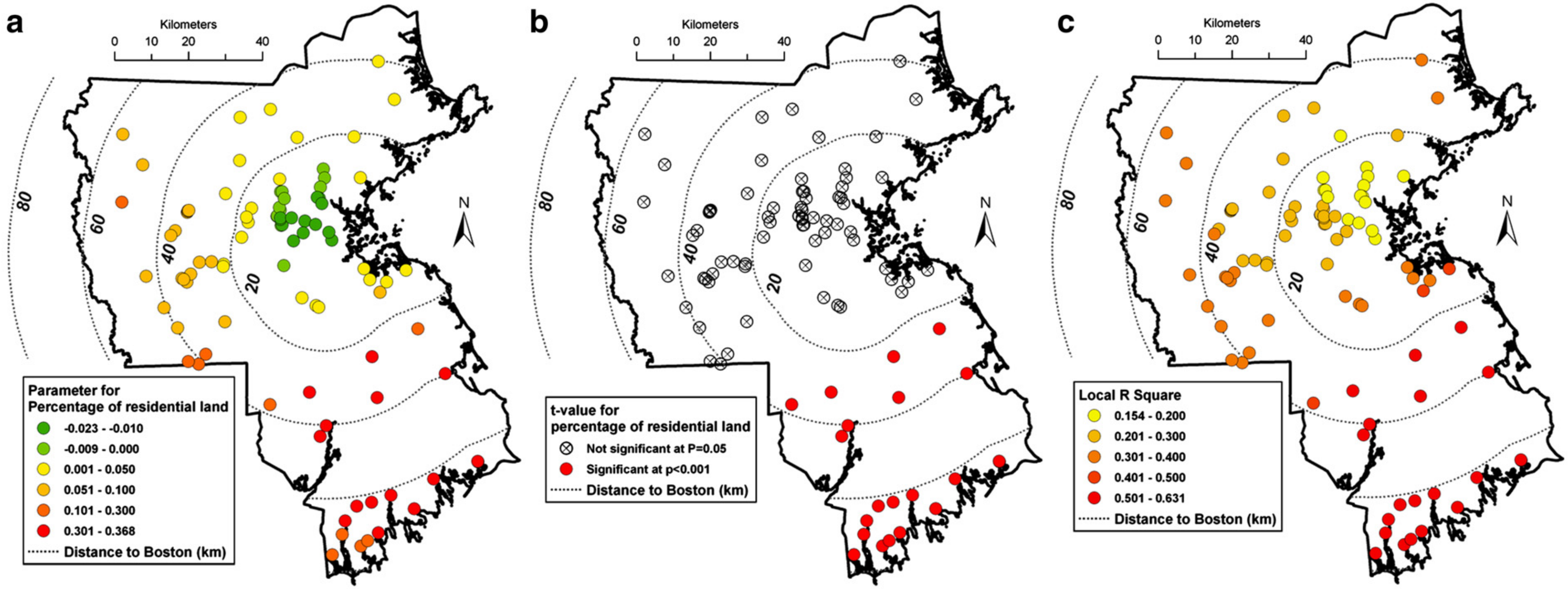

Fig. 9 – Results of the GWR model for SC and percentage of residential land: (a) parameter estimates for percentage of residential land; (b) t-value for percentage of residential land; (c) local R^2.

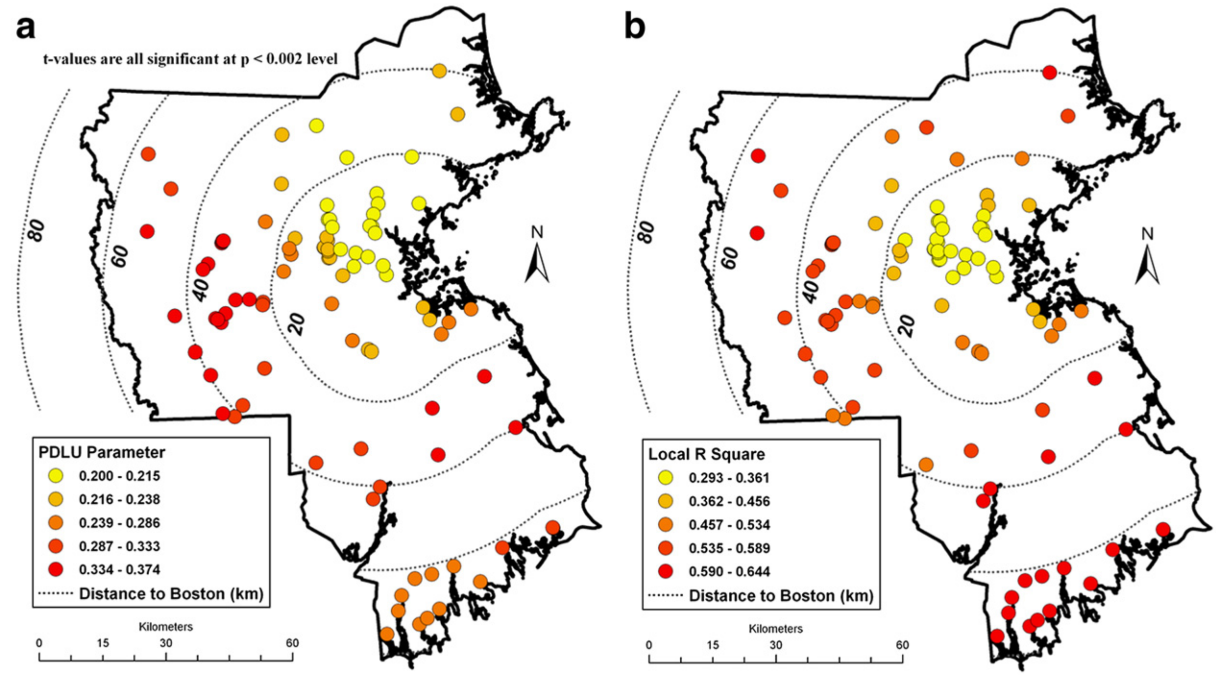

Fig. 10 – Results of the GWR model for SC and percentage of developed land: (a) parameter estimates for percentage of developed land; (b) local R^2.

a contribution to SC in the surface water, and therefore agriculture activities including application of fertilize and livestock are more important pollution sources than they are in other parts of the study area. As shown in the map of *t*-test on the parameters, most of the correlations are significant (Fig. 3b). The local R^2 also vary over space (Fig. 3c). High R^2 values appear in the middle part of the area between 20-km and 40-km buffers, where the concentration of SC is positively related to the percentage of agricultural land, and the percentage of agricultural land can explain more than 70% of the spatial variation in SC. Low R^2 values are observed in the south part of the study area, where the percentage of agricultural land can explain down to only 30% of the variance in SC. However, it is still higher than that obtained from the OLS model (R^2=0.134, Table 2).

The maps of parameter estimates, *t*-test, and local R^2 from the GWR model for SC and forest land are shown in Fig. 4. Based on the result of OLS model, percentage of forest land is a good predictor of water quality. Percentage of forest land has a significant negative correlation with the concentration of SC (r=−0.646, p<0.001; Table 1), indicating that higher percentage of forest land is related to better water quality. Similarly, negative correlations are also observed for almost all the study area except one sampling site from the GWR model (Fig. 4a). However, the correlation is not significant at most of sites within metropolitan Boston area (Fig. 4b), suggesting that forest land is a less important predictor of water quality in highly-urbanized area than in less-urbanized area. The local R^2 values for SC and forest land also exhibit a huge spatial variability, ranging from 0.136 to 0.989 (Fig. 4c).

Open undeveloped land is the weakest predictor of water quality in the OLS model. It has no significant correlations with any water quality indicators. However, it has significant positive local correlations with SC in the western part of the area outside of the 20-km buffer identified by GWR (Fig. 5a and b). The local R^2 values for SC and open undeveloped land also exhibit a great spatial variability, ranging from 0.184 to 0.839 (Fig. 5c). The result indicates that percentage of open undeveloped land also becomes a good predictor in some areas, especially the western part of the area outside of the 20-km buffer, which is mainly rural.

Fig. 6 shows the maps of parameter estimates, *t*-test, and local R^2 from the GWR model for SC and commercial land. Commercial land has a significant positive correlation with SC in the OLS model (r=0.694, p<0.001; Table 1). Significant positive local correlations are still observed for most of the sample sites. However, a clear spatial non-stationarity exists as identified from the results of the GWR model. More significant correlations and higher local R^2 values are found for the sampling sites within the areas outside of the 20-km buffer, mainly rural areas. Less significant correlations and lower local R^2 values are located inside of the 20-km buffer and the northern part of areas between 20-km and 40-km buffers, which are mainly suburban and central city areas. Percentage of commercial land can explain more than 90% of the spatial variation in SC in rural areas but lower than 30% in central city areas (Fig. 6c). This result indicates that commercial land has more important impact on the concentration of SC in less-urbanized areas than it in highly-urbanized areas. Same amount of increase in the percentage of commercial land may cause more increase in the concentration of SC in less-urbanized areas than in highly-urbanized areas.

Percentage of industrial land is a good predictor of SC as shown in the result of OLS model. High percentage of industrial land is related to higher concentration of SC (r=0.649, p<0.001; Table 1). Similarly, positive correlations are found for most of sampling sites by the GWR model (Fig. 7a).

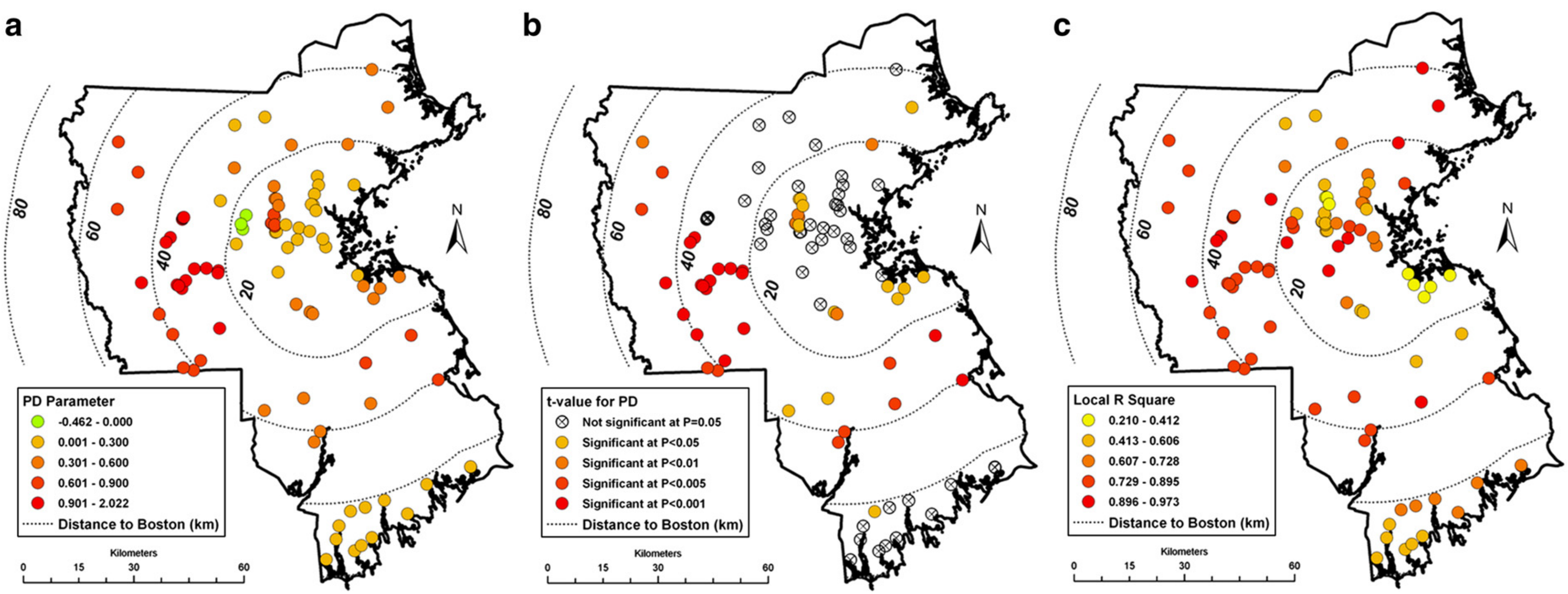

Fig. 11 – Results of the GWR model for SC and population density: (a) parameter estimates for population density; (b) *t*-value for population density; (c) local R^2.

However, a lot of sites within 20-km buffer, northern part of the areas between 20-km and 40-km buffers, and the most southern part of the study area have no significant correlations (Fig. 7b). A great spatial variability also exists in local R^2 values, ranging from 0.081 to 0.979 (Fig. 7c).

Fig. 8 shows the maps of parameter estimates, *t*-test, and local R^2 from the GWR model for SC and recreation use. A significant positive correlation is found between SC and percentage of recreation use by the OLS model ($r=0.326$, $p=0.001$; Table 1). However, both negative and positive correlations are observed in the study area by the GWR model, and a clear spatial pattern can be identified from the maps (Fig. 8a and b). Most sites inside the 20-km buffer from Boston, mainly metropolitan area, have significant or non-significant negative correlations, while most of the sites outside the 20-km buffer have positive correlations. This result suggests that higher percentage of land for recreation use is related to good water quality in metropolitan area, while recreation activities in rural area may contribute to water pollution. A great spatial variability also exists in local R^2, ranging from 0.120 to 0.948 (Fig. 8c), and every local R^2 is higher than the R^2 in the OLS model ($R^2=0.106$, Table 2). This example confirms again that GWR model not only reveal the different roles played by the same type of land use in different places but also improve the ability of model to explain local situation of water quality.

As opposed to the significant positive correlation between percentage of residential land and SC concentration found by the OLS model ($r=0.403$, $p<0.001$; Table 1), both positive and negative correlations are observed by the GWR model (Fig. 9a and b). However, all the negative correlations are not significant, and only the southern part of the area outside of 20-km buffer shows significant positive correlations. Percentage of residential land can explain more than 60% of the spatial variation in the concentration of SC (Fig. 9c). This area was identified as “sprawl frontier” under high development pressures by Massachusetts Audubon Society because of its fast new single-family housing development but relatively low developed (Mass Audubon, 2003). The result from GWR model in this study confirms this point. Residential land plays a more important role in affecting water quality in this “sprawl frontier” than that in the other areas.

Similar to the result of the OLS model for the correlation between PDLU and SC ($r=0.680$, $p<0.001$; Table 1), significant positive correlations of these two variables are also found for all the sampling sites by the GWR model (Fig. 10a). However, a considerable spatial non-stationarity and a clear spatial pattern can be also identified from the maps of parameter estimates and local R^2 (Fig. 10). Percentage of developed land has stronger relationships with SC concentration in areas outside of 20-km buffer than that inside, indicating that new development including residential, industrial, and commercial land in less-urbanized area will contribute more to the increase of SC than that in highly-urbanized area.

Another composite land use indicator, population density, shows a significant positive correlation with SC in the OLS model ($r=0.598$, $p<0.001$; Table 1). However, based on the results of the GWR model, it only shows significant positive correlations with SC at some sites within rural and suburban areas. Most of sites in highly-urbanized area have no significant correlations. Similar to commercial, industrial lands and recreation use, PD also has the highest local R^2 in the western part of the area outside the 20-km buffer, which is another “sprawl frontier” identified by Massachusetts Audubon Society (Mass Audubon, 2003). This result confirms again that new development plays a more important role in affecting water quality in the areas with fast development rate but low percentage of developed land.

5. Conclusions

This study examined the relationships between nine land use and fourteen water quality indicators using both OLS and GWR models. OLS found that most of water quality indicators, including SC, dissolved ions and solid, and dissolved nitrogen parameters, have significant correlations with most of land use indicators, including population density, percentages of developed land, agricultural, forest, commercial, industrial, and residential lands.

However, Most GWR models show great improvements of model performance over their corresponding OLS models, which is proved by *F*-test and the comparisons of model R^2 and AIC_c from both GWR and OLS. A dramatic improvement in R^2 of GWR over OLS is observed for every pair of water quality and land use indicators.

Many GWR models also successfully reduce spatial autocorrelations examined by Moran's I statistics. Significant spatial autocorrelations are found in all the OLS models for SC, Ca, Na, and Cl and some OLS models for other water quality indicators, as the OLS models are not able to account for the spatial dependences in water quality. The spatial autocorrelations reduce the efficiency of the regression and make the OLS models unsuitable for identifying the relationships between land use and water quality indicators. Fortunately, GWR models improve the reliabilities of the relationships by reducing the spatial autocorrelations in residuals.

The most novel and interesting finding of this study is that the relationships of land use and water quality are not constant over space, and GWR technique provides a simple but powerful tool to explore the spatially varying relationships. The visualization of the GWR model local parameter estimates, *t*-test, and local R^2 highlight the great spatial variations in the impacts of different land use types on different water quality indicators and help identify their spatial patterns. GWR models also reveal previously unknown information on the different roles of a land use type plays in different parts of the study area, and so improve the model ability to explain local situation of water quality. For example, the OLS model finds a significant negative correlation between percentage of agricultural land and concentration of SC for the study area, while the corresponding GWR model reveals that both significant negative and positive correlations exist in this area. In highly-urbanized area, higher percentage of agricultural land is related to lower SC concentration, because it means lower percentage of developed land including industrial, commercial, and residential lands, which might be the major pollution sources there. In contrast, in some sampling sites within less-urbanized area, higher percentage of agricultural land is related to higher concentration of SC, as agricultural activities become important water pollution sources.

The results of this study suggest that GWR technique can serve as a useful tool for environmental research and management at watershed, regional, national and even global scales. Environmental scientists and environmental protection agencies are all concerned with how natural environment affected by different natural and anthropogenic factors, such as soil, climate, land use, human activities, and policy. All these factors are changing over space. They are different in different watersheds, regions, and countries, and may play different roles in natural environment in different areas. As a result, spatial non-stationarity may exist in their relationships with environment, which can be examined by GWR models. Furthermore, local environmental protection agencies and decision makers are also concerned with the situation and causes of environment pollution in their local places. After the local situation and problem are revealed using GWR, they may be able to adopt appropriate policies and techniques suitable to the local environment.

REFERENCES

Anselin L, Syabri I, Kho Y. GeoDa: an introduction to spatial data analysis. Geogr Anal 2006;38:5–22.

Basnyat P, Teeter LD, Flynn KM, Lockaby BG. Relationships between landscape characteristics and nonpoint source pollution inputs to coastal estuaries. Environ Manag 1999;23:539–49.

Brunsdon C, Fotheringham S, Charlton M. Geographically weighted regression – modeling spatial non-stationarity. Statistician 1998;47:431–43.

Clark SD. Estimating local car ownership models. J Transp Geogr 2007;15:184–97.

Dow CL, Zampella RA. Specific conductance and pH as indicators of watershed disturbance in streams of the New Jersey Pinelands, USA. Environ Manag 2000;26:437–45.

Farrow A, Larrea C, Hyman G, Lema G. Exploring the spatial variation of food poverty in Ecuador. Food Policy 2005;30:510–31.

Fishman MJ, Friedman LC. Methods for determination of inorganic substances in water and fluvial sediments. U.S. Geological Survey Techniques of Water-Resources Investigations; 1989. Book 5, Chapter A1, Available online at http://water.usgs.gov/pubs/twri/. (Accessed on August 21, 2007).

Fotheringham AS, Charlton ME, Brunsdon C. Spatial variations in school performance: a local analysis using geographically weighted regression. Geogr Environ Model 2001;5:43–66.

Fotheringham AS, Brunsdon C, Charlton M. Geographically weighted regression: the analysis of spatially varying relationships. Chichester: Wiley; 2002. 269 pp.

Friedman LC, Erdmann DE. Quality assurance practices for the chemical and biological analyses of water and fluvial sediments. U.S. Geological Survey Techniques of Water-Resources Investigations; 1982. Book 5, Chapter A6, Available online at http://water.usgs.gov/pubs/twri/. (Accessed on August 21, 2007).

Hamilton LC. Regression with graphics: a second course in applied statistics. Belmont, California: Duxbury Press; 1992. 363 pp.

Harries K. Extreme spatial variations in crime density in Baltimore County, MD. Geoforum 2006;37:404–6.

Ishizawa H, Stevens G. Non-English language neighborhoods in Chicago, Illinois: 2000. Soc Sci Res 2007;36:1042–64.

Jarvie HP, Oguchi T, Neal C. Exploring the linkages between river water chemistry and watershed characteristics using GIS-based catchment and locality analyses. Reg Environ change 2002;3:36–50.

Little JL, Saffran KA, Fent L. Land use and water quality relationships in the lower Little Bow River Watershed, Alberta, Canada. Water Qual Res J Can 2003;38:563–84.

Longley PA, Tobón C. Spatial dependence and heterogeneity in patterns of hardship: an intra-urban analysis. Ann Assoc Am Geogr 2004;94:503–19.

Mass Audubon. Losing Ground: At What Cost. Massachusetts Audubon Society; 2003. Available at www.massaudubon.org/losingground. (Accessed on July 11, 2007).

MassGIS. Massachusetts land use datalayers and land use Summary Statistics; 2005. Available at http://www.mass.gov/mgis/landuse_stats.htm; http://www.mass.gov/mgis/lus.htm (Accessed on September 19, 2005).

Mehaffey MH, Nash MS, Wade TG, Ebert DW, Jones KB, Rager A. Linking land cover and water quality in New York City's water supply watersheds. Environ Monit Assess 2005;107:29–44.

Omernik JM. Ecoregions of the conterminous United States. Ann Assoc Am Geogr 1987;77:118–25.

Rodriguez W, August PV, Wang Y, Paul JF, Gold A, Rubinstein N. Empirical relationships between land use/cover and estuarine condition in the Northeastern United States. Landsc Ecol 2007;22:403–17.

Shi H, Laurent EJ, LeBouton J, Racevskis L, Hall KR, Donovan M, et al. Local spatial modeling of white-tailed deer distribution. Ecol Model 2006;190:171–89.

Simcox AC. Water resources of Massachusetts. USGS water-resources investigations report 90-4144. U.S. Geological Survey; 1992. 94 pp. Available from: http://infotrek.er.usgs.gov/pubs/. (Accessed on December 19, 2005).

Tong STY, Chen W. Modeling the relationship between land use and surface water quality. J Environ Manag 2002;66:377–93.

Tu J, Xia ZG. Assessing the impact of land use changes on water quality across multiple spatial scales in eastern Massachusetts. Middle States Geogr 2006;39:34–42.

Tu, J., Xia, Z.G., Examining spatially varying relationships between land use and water quality using Geographically Weighted Regression II: results interpretation and implication. Sci Total Environ, submitted for publication.

Tu J, Xia ZG, Clarke KC, Frei A. Impact of urban sprawl on water quality in eastern Massachusetts, USA. Environ Manag 2007;40:183–200.

U.S. Geological Survey, National field manual for the collection of water-quality data: U.S. Geological Survey Techniques of Water-Resources Investigations, Book 9, Chaps. A1-A9, various dates, ranging from Oct. 1997 to Aug. 2007. Available online at http://pubs.water.usgs.gov/twri9A. (Accessed on August 21, 2007).

Wang X, Yin ZY. Using GIS to assess the relationship between land use and water quality at a watershed level. Environ Int 1997;23:103–14.

Wang Q, Ni J, Tenhunen J. Application of a geographically-weighted regression analysis to estimate net primary production of Chinese forest ecosystems. Glob Ecol Biogeogr 2005;14:379–93.

Williams M, Hopkinson C, Rastetter E, Vallino J, Claessens L. Relationships of land use and stream solute concentrations in the Ipswich River basin, Northeastern Massachusetts. Water Air Soil Pollut 2005;161:55–74.

Woli KP, Nagumo T, Kuramochi K, Hatano R. Evaluating river water quality through land use analysis and N budget approaches in liverstock farming areas. Sci Total Environ 2004;329:61–74.

Xiao H, Ji W. Relating landscape characteristics to non-point source pollution in mine waste-located watersheds using geospatial techniques. J Environ Manag 2007;82:111–9.

Yu D. Spatially varying development mechanisms in the greater Beijing area: a geographically weighted regression investigation. Ann Reg Sci 2006;40:173–90.

Zhang L, Bi H, Cheng P, Davis CJ. Modeling spatial variation in tree diameter-height relationships. For Ecol Manag 2004;189:317–29.