Chia-Rung Yang, Hong Xia

Master's Candidate in Geographic Information Science for Development and Environment

Department of International Development, Community and Development

Clark University

Contact Information:

Chia-Rung Yang
chyang@clarku.edu
1-617-775-3206
950 Main Street, IDCE Dept.
Worcester, MA 01610-1477

Hong Xia
hxia@clarku.edu
1-508-762-7591
950 Main Street, IDCE Dept.
Worcester, MA 01610-1477

Characterizing Factors Associated with Low Birth Weight in Massachusetts Using Geographically Weighted Logistic Regression

Abstract

Low birth weight (LBW) is closely related with negative effects on neonatal health development. In this study, we examined the factors, both maternal and environmental, that might contribute to LBW. The study population consisted of 623,844 births from 2000 to 2007 in Massachusetts. Infants with a birth weight below 2,500 grams were classified as LBW. The related birth data was obtained from the Massachusetts Department of Public Health (MassDPH). The data on exposure to environmental pollution, specifically lead in our study, was from the US Environmental Protection Agency (EPA) Risk-Screening Environmental Indicators (RSEI) Model. We selected variables to be incorporated in the analysis based on their calculated unadjusted odds ratio. Geographically weighted logistic regression (GWLR) was then applied in our analysis to predict the likelihood of LBW cases. The advantage of GWLR, compared with methods applied in previous studies, is that it takes into account spatial autocorrelation and spatial non-stationarity during the process. Our findings indicated that LBW is highly associated with the infant's plurality and infant's gestational age, whereas the association between LBW and other variables such as mother's age, smoking history and exposure to lead, etc. varied across the entire Massachusetts. The analysis could be further improved by carrying out a cross validation on each GWLR model to get a better idea of how the model performs.

1 Introduction

1.1 Background

Low birth weight (LBW), as has been proven in various studies, is closely related with fetal and neonatal mortality and morbidity, inhibited growth and cognitive development, chronic diseases later in life (World Health Organization, 2010). The factors related to LBW can be biological, socioeconomical and environmental. Among environmental factors, maternal exposure to toxic chemicals, especially air pollutants, has been explored in various studies on LBW. It is generally agreed that air pollution has negative effects on infant's birth weight. In this study, we tried to integrate air pollution and other socioeconomic factors into a regression analysis and find out each factor's association with LBW. We explored the association between LBW and human exposure to lead, as well as characterized the effect from other related variables. Particularly, geographically weighted logistic regression (GWLR) was applied to incorporate non-numerical variables in the model and characterize the spatial non-stationarity of the association.

1.2 Literature review

1.2.1 Low Birth Weight and Its Factors

Low birth weight (LBW) is often defined as the birth weight less than 2,500 (Ashdown-Lambert, 2005). Various studies tried to find the association between LBW and its factors. Three types of factors have been examined in previous studies:

Biological factors: biological plausibility is often referred when researchers tried to verify the causative relationship between a factor and LBW. These biological mechanisms have been summarized by Kannan et al. (2007). Various biological factors, such as diseases or special health

status occurring with LBW, have been indicated in various studies. The connection between diseases or health and LBW can be explained by either maternal nutrition (Barker, 1995) or genetically determined physiological status (Hattersley & Tooke, 1999).

Socioeconomic factors: income, ethnicity, education and marital status have been suggested to associate with LBW as well (Krieger, 1992; Pattenden et al. 1999; Ash et al., 2004; Cramer, 1995). Previous studies have found that these socioeconomic statuses often relate to human health status and access to medical care (Krieger, 1992). In addition, these socioeconomic statuses are often related to behaviors such as smoking, alcohol and drug use, as well as psychological statuses such as stress (Cramer, 1995).

Environmental factors: the effect of environmental factors has been examined in various ways. Dozens of studies have explored the association between environmental toxic chemicals and LBW. Most of them were conducted on the association with air pollutants (Shah et al., 2011), while a few of them found the connection with water contamination as well (Witkowski and Johnson, 1992; Bove et al., 1995; Villanueva et al., 2004).

1.2.2 Air Pollution and LBW

Through reviewing 17 case studies that explored the association between LBW and air pollution we identified chemicals that have been explored in previous studies, as well as the ways that previous studies have applied to estimate human exposure to air pollutants.

Six air pollutants listed on the U.S. Environmental Protection Agency (EPA) National Ambient Air Quality Standards (NAAQS) have been frequently explored in previous studies: carbon monoxide (CO), lead (Pb), nitrogen dioxide (NO_2), particle pollution (PM_{10}, $PM_{2.5}$), ozone (O_3) and sulfur dioxide (SO_2). These chemicals are considered harmful to public health and the environment (US Environmental Protection Agency, 2011). Slama et al. (2007) found that high levels of $PM_{2.5}$ and $PM_{2.5}$ absorption would result in LBW. Dugandzic et al. (2006) examined the association between exposure

to ambient air pollution and LBW, and they found that high exposures to SO_2 and PM_{10} during the first trimester suggested an increased risk of delivering a LBW infant. Ha et al. (2001) has also carried out similar research in Seoul. CO, NO_2, SO_2 and total suspended particle concentrations were found to be risk factors for LBW in the first trimester of pregnancy period. In addition to these chemicals, Aguilera et al. (2009) discovered apparent association between benzene, toluene, ethylbenzene, and xylenes (BTEX) and reduced birth weight, thus reflecting the negative role of vehicle exhaust pollutants in reproductive health. These studies examined the association between LBW and various chemicals. While various conclusions were drawn in different studies, most of them suggested that the impact of air pollution on LBW cannot be ignored and more specific research needs to be done to further identify its impact.

In the studies on associations between air pollution and health outcomes, a common problem is to find an appropriate method to estimate human exposure to pollutants based on limited number of air pollution monitoring stations. These stations can only provide observed values at limited points. In the studies we reviewed, three approaches were applied to estimate human exposure to air pollutants.

The first approach is based on proximity. Each case's exposure to air pollution is assigned a value from its closest monitoring station. Wilhelm and Ritz (2005) classified cases according to their proximity to monitoring stations. Considering only the proximity to pollution source, however, could probably cause the misclassification of exposure and raise biases (Ryan et al., 2007a).

The second approach interpolates the chemical concentration at specific locations based on the data from a monitoring network. An assumption of this approach is that the change of pollutant concentration between two monitoring sites can be fitted into a mathematical function. Inverse distance weight (IDW) was employed in several studies. Chan et al. (2009) used the Kriging method to estimate concentration, as well as generalized additive model (GAM) to estimate the effect of increasing

pollutant exposure on asthma outpatient cases. Yanosky et al. (2008) also employed GAMs to circumvent the limitations of using community monitoring data and other simple spatial interpolation.

The third approach is to model the distribution of pollutants based on information other than monitoring data. Slama et al. (2007) applied land use regression (LUR) to model concentrations of traffic-related atmospheric pollutants based on land use types. This model assumed that there is certain association between land use type and air pollutant concentration. Aguilera et al. (2009) employed LUR with four variables (altitude, land coverage, and two road length indicators) to estimate the intra-urban variation of pregnant women's exposure to air pollutants. Overall, LUR usually performs better on showing intra-urban and local variations than inter-urban variations (Ryan et al., 2007b). Other models were also applied for different scenarios. Gryparis (2007) employed latent variable semiparametric regression models in the greater Boston area. McEntee and Ogneva-Himmelberger (2008) employed the US EPA's National Air Toxics Assessment (NATA) data, derived from Hazardous Air Pollutant Exposure Model (HAPEM), to identify the spatial hot spots of elevated diesel particular matter (DPM). Yu et al. (2009) applied Bayesian maximum entropy (BME) to model residential exposure to pollutants and found that it performs better than Kriging.

Comparing the advantages of LUR, IDW interpolation and GAM, Brauer et al. (2008) found that LUR models can better predict personal exposure than IDW interpolation, although it might also incorporate more misclassification than using the observation at the closest monitoring site. Hart et al. (2009) also found that GAMs performed better than IDW interpolation in their case. However, using the Kriging models with different algorithms might contribute to different results (Liao et al. 2006). Overall, the exposure modeling can be improved under various circumstances and hybrid methods might be more suitable to overcome specific problems (Jerrett et al., 2005).

1.2.3 Exploration of Association

Depending on the types of variables (numerical, binary, ordinal, categorical, etc.), various statistical methods have been employed to evaluate the relationship between LBW and exposure to air pollutants. Among these multivariate statistical methods, linear regression has been applied in the studies by Cramer et al. (1995), Ash et al. (2002), Basu et al. (2004), and Aguilera et al. (2009), which only involve numerical variables. In these studies, the dependent variable birth weight was analyzed as a numerical variable and measured by grams, so were all the independent variables such as exposure to air pollution, household income, percentage of specific ethnicity, percentage of electoral voting, etc.

In some studies (Lee et al., 2003; Gouveia et al., 2004), instead, birth weight was defined as a binary variable that only contains the value TRUE or FALSE. 2,500 grams were used as the breakpoint to distinguish LBW cases from those which are not. Logistic regression was applied in these studies to deal with binary variables. Adjusted odds ratios between LBW and independent variables can be calculated from the regression coefficients.

Linear and logistic regressions are more commonly applied than other regression models. However, in one of the studies we reviewed, Slama et al. (2007) used Poisson regression to estimate the prevalence ratios (PR) of LBW due to traffic related atmospheric pollutants.

1.2.4 Geographic Characteristics in Analysis

We also examined how geographic characteristics were addressed in previous studies. We reviewed these studies from three perspectives: the definition of study area, the method to locate cases, and whether or not geographic locations were incorporated into their multivariate statistics.

Studies on the association between LBW and air pollution often incorporate the air pollution monitoring data at specific locations. The definition of study area largely relies on the availability of monitoring data. Among the ten studies using air monitoring data that we reviewed (Table 1), six were conducted within urban areas and only one was conducted at the state and province level (Dugandzic et

al., 2006). Most of them utilized the existing monitoring data within study areas. Generally those monitoring data have a more even and dense coverage within cities. In rural areas, instead, the poor coverage of monitoring network makes it much more difficult to model the exposure at specific locations.

Table 1 The study areas of ten studies that involve monitoring data

Level of Study Area	Studies
National	US (Liao et al., 2006)
State and Province	Nova Scotia, Canada (Dugandzic et al., 2006)
County	Shelby County, Tennessee, US (Ozdenerol, 2005) San Diego County, California, US (Ross et al., 2006)
Urban Area	Six Cities in the Northeastern US (Maisonet et al., 2001) Sao Paulo, Brazil (Gouveia et al., 2004) Munich, Germany (Slama et al., 2007) Vancouver, British Columbia, Canada (Nethery et al., 2008) Vancouver, British Columbia, Canada (Brauer et al., 2008) Sabadell City, Spain (Aguilera et al., 2009)

Within each study area, researchers positioned case locations by various ways depending on the resolution of residential data. Geocoding is often used to find the geographic coordinates of each case based on precise street addresses. When precise street addresses are not available, administrative areas such as urban areas (Gouveia, 2004), census areas (Pattenden, 1999) and ZIP code areas (Brauer, 2008) were used to locate birth cases. In addition to identifying geographic location of individuals, some ecological studies utilized aggregated regional data to explore this association (Pattenden, 1999; Ash, 2004). These studies show their strength to classify cases properly, while more or less resulting in ecological fallacy (Pattenden, 1999).

Even though most studies located residential addresses to estimate human exposure to air pollution, a few of them incorporated geographic locations as an independent variable in their multivariate analyses. Brauer et al. (2008) divided cases according to their proximities to roads. Physical distance was employed as a proxy of exposure to traffic-related air pollution. Two physical

distances, 50 meters and 150 meters, were used to classify the cases into three groups. Proximity was employed as a categorical variable. Their study characterized the geographic heterogeneity of the association. Aside from this, all the studies we reviewed did not model the geographic heterogeneity of association.

1.3 Research Objective

This study is aimed at building a model with variables including level of exposure to air pollution and examining each factor's association with LBW. Among six air pollutants listed on the EPA's NAAQS we chose lead. Lead is proved to have negative effects on infant's birth weight and children's health development later on, yet the exposure to lead in the air seems to be rarely studied. In our study, logistic regression was employed for two reasons. First, logistic regression is able to estimate the adjusted odds ratio of each variable. This ratio shows the variable's association with LBW while excluding the effects of other variables. Second, our dataset contains many non-numerical variables. A linear model is not able to incorporate these variables into the regression, while logistic regression can properly deal with these variables.

In addition, this study covers the entire state of Massachusetts. While most of the previous studies only covered urban areas, this study includes the birth cases in the entire state. For such a large study area, a global regression model is not able to characterize the spatial heterogeneity of these associations. In order to overcome the shortcomings of global models, based on the concept of geographically weighted regression (GWR) (Fotheringham et al., 2002) and its generalized variations, we applied geographically weight logistic regression (GWLR) in this study.

2 Methods

2.1 Data

The LBW data, provided by the Massachusetts Department of Public Health (MassDPH), recorded 623,844 births from 2000 to 2007 with variables including infant's birth weight, sex and plurality as well as the mother's age, marital status, race, education, gestational age, smoking history and other health conditions. The variable types include numerical, ordinal and binary. The geographic locations of all birth cases were geocoded by the MassDPH. Geographical location of individual birth is available at the census block level (i.e. no street address was available).

Through a preliminary examination of the data, we have excluded the records with incomplete information, including a missing part in southeastern Massachusetts in 2007 due to geocoding failures. Among the recorded 623,844 births, 584,603 (93.7%) with complete information were included in our analysis. Among these cases, the distribution of LBW is shown in Figure 1.

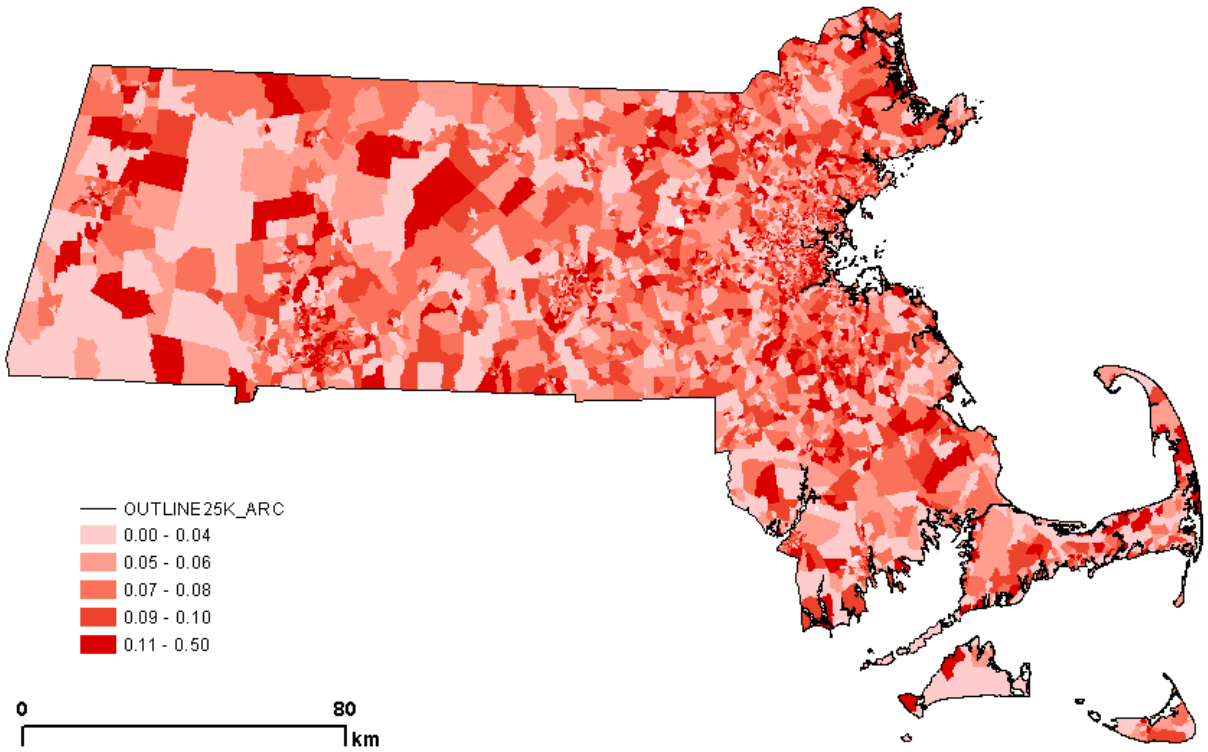


Figure 1 The spatial distribution of the LBW cases

We selected 22 variables from the dataset based on odds ratios. Odds ratio is an index of association between two binary variables. For any binary variable, all cases can be split into two groups according to their values of that variable (i.e. TRUE or FALSE). In this study, the odds ratio is the ratio of the odds of LBW occurring in one group to the odds of LBW occurring in another group. This value indicates if a variable is positively or negatively associated with LBW. For example, the odds ratio of mother's race and LBW is the odds of LBW occurring in the non-white ethnicities to the odds of LBW occurring in among the white. If the odds ratio of a variable and LBW is larger than 1.0, this variable is positively associated with LBW; if it is smaller than 1.0, it is negatively associated with LBW. Any variable with an odds ratio deviating from 1.0 was selected to be incorporated in the logistic regression model.

These 22 variables selected from our dataset are: plurality (PLUR), mother's age (MAGE), marital status (MARSTST), mother's race (MRACE), clinical estimate of gestational age (CEGA, in weeks), Kessner index (KESS, categorical variable used to indicate the adequacy of mother's care during pregnancy), number of cigarettes during pregnancy (CIGDP), maternal weight gained or lost (MWGL, in grams), cardiac disease, gestational diabetes, eclampsia, hydramnios/oligohydramnios, hemoglobinopathy, chronic hypertension, pregnancy-related hpyertension, incomplete cervix, lupus erythematosus, previous infant 4,000+ grams, previous preterm infant, renal disease, sickle cell disease, and uterine bleeding. The last fourteen variables are all related to mother's health condition. They were aggregated to a new variable OTHER, which indicates if they had any of these fourteen health conditions. If any of these fourteen variables has the value TRUE, the value of OTHER will be TRUE as well. Other variables were converted to binary format as described in Table 2.

Originally there were four numerical variables among the 22 selected variables: mother's age (MAGE), clinical estimate of gestational age (CEGA), number of cigarettes during pregnancy (CIGDP), maternal weight gained or lost (MWGL, in grams). When incorporating a numerical variable

into a logistic regression model, it is assumed that those cases with higher value of this numerical variable are more likely to be in one category of the binary dependent variable, whereas those cases with lower value of this numerical variable are more likely to be in another category of the binary dependent variable. In our dataset, however, we observed that both the younger mother group and the elder mother group yield a higher ratio of LBW. The same phenomenon was observed in the variations of CEGA. Apparently their relationships with LBW are not appropriate to be described with only one numerical variable in a logistic regression model. Therefore, in order to be properly incorporated into the model, MAGE was converted to two binary variables: MAGE17, which shows if the mother's age is equal to or younger than 17, and MAGE 35, which indicates if the mother's age is equal to or elder than 35. The former indicates if a mother is a teenager, while the latter indicates if a mother was pregnant over the age of 35, which increases the risk of several infant's health conditions (Southern California Center for Reproductive Medicine, 2011).

In the same way, CEGA was converted to two binary variables: CEGA 36 and CEAG 43, which refer to the gestational ages equal to or smaller than 36 and equal to or greater than 43 respectively. These two numbers were determined based on the definitions of preterm birth (World Health Organization, 1976) and postmature birth (Mosby, 2009). CIGDP was converted to a binary variable in order to indicate whether or not a mother smoked during pregnancy. Only MWGL, whose odds (i.e. chance of happening) are proportional to the ratio of LBW, was kept as a numerical variable.

After conversion, the odds ratios of these new binary variables were calculated again. The distribution and odds ratio of each binary variable is shown in Table 3. CEGA 43 was excluded later since there are only 8 LBW cases with gestational age over 43 weeks.

Table 2 Value Conversion for Variables

Variable	Value Conversion	Included in Model
PLUR	0-one child, 1-two or more children	Yes
MAGE17	If MAGE $\leq$ 17, 0-no, 1-yes	Yes

MAGE35	If MAGE ≥ 35, 0-no, 1-yes	Yes
MARSTST	0-married, 1-not presently married	Yes
MRACE	0-white, 1-not white	Yes
CEGA36	If CEGA ≤ 36, 0-no, 1-yes	Yes
CEGA43	If CEGA ≥ 43, 0-no, 1-yes	No
KESS	0-adequate, 1-inadequate, intermediate and others	Yes
CIGDP	0-no, 1-yes	Yes
MWGL	Not convert (numerical data)	Yes

Table 3 The distribution and pivot with LBW of 10 binary variables

Variable	Value	Distribution		Pivot with LBW			Crude Odds Ratio	Adjusted Odds Ratio
		Sum	Percentage	FALSE	TRUE	Percentage		
PLUR	0	558137	95.47%	529306	28831	5.17%	-	-
	1	26466	4.53%	12367	14099	53.27%	20.93	0.094
MARSTST	0	415999	71.16%	388181	27818	6.69%	-	-
	1	168604	28.84%	153492	15112	8.96%	1.37	0.708
MRACE	0	424758	72.66%	396214	28544	6.72%	-	-
	1	159845	27.34%	145459	14386	9.00%	1.37	0.668
KESS	0	463243	79.24%	430819	32424	7.00%	-	-
	1	121360	20.76%	110854	10506	8.66%	1.26	1.006
CIGDP	0	537434	91.93%	499740	37694	7.01%	-	-
	1	47169	8.07%	41933	5236	11.10%	1.66	0.498
MAGE17	0	572855	97.99%	531153	41702	7.28%	-	-
	1	11748	2.01%	10520	1228	10.45%	1.49	0.696
MAGE35	0	451090	77.16%	419113	31977	7.09%	-	-
	1	133513	22.84%	122560	10953	8.20%	1.17	0.996
CEGA36	0	535772	91.65%	522060	13712	2.56%	-	-
	1	48831	8.35%	19613	29218	59.83%	56.72	0.028
CEGA43	0	584330	99.95%	541408	42922	7.35%	-	-
	1	273	0.05%	265	8	2.93%	0.38	N/A*
OTHER	0	507044	86.73%	476719	30325	5.98%	-	-
	1	77559	13.27%	64954	12605	16.25%	3.05	0.542

* This variable was not incorporated when calculating adjusted odds ratios.

The finest geographic scale, census block, was used to locate each case. We assigned a pair of coordinates to each case using the centroid of the block in which the case is located. The geographic locations were employed for two purposes: to generate weight files that define weights based on

distance and to find the correspondent hazard scores, which are addressed in next chapter and this chapter respectively.

The data on exposure to lead are from the US Environmental Protection Agency (EPA) Risk-Screening Environmental Indicators (RSEI) Model. It provided data on the relative levels of exposure in grid format (cell resolution is 810*810m). This model estimates people's exposure to chemicals based on the total release amount from the EPA Toxics Release Inventory (TRI) sites. Three results of 591 chemicals are all available: the pound-based, hazard-based and risk-related scores (Table 4). The hazard-related score (HAZARD), which does not involve the effects of demographic data, was employed in this study.

Table 4 The Results of EPA RSEI

Results	Calculation
Risk-related results	Surrogate Dose x Toxicity Weight x Exposed Population
Hazard-based results	Pounds x Toxicity Weight
Pounds-based results	TRI Pounds released

The exposure data were provided as raster files. We overlaid these raster files with the vector file of census blocks. If the area of a block is larger than one cell, the average exposure value of its overlapping cells was assigned to that block. If the area of a block is smaller than a cell, the exposure of that block was obtained by overlapping the block centroid with the raster file of exposure.

In total, eleven variables were included: all variables in Table 2 except CEGA43, plus HAZARD and OTHER. Only HAZARD and MWGL are numerical variables. The other nine variables are binary.

2.2 Method

In traditional statistics, one of the most commonly used ways to explore the relationship between dependent and independent variables is the ordinary least squared (OLS) regression. One of its core assumptions is that the errors for each observation are randomly distributed. However, most of the spatial data violates this assumption due to spatial autocorrelation. The OLS model is not able to characterize spatial non-stationarity, either. Geographically weighted regression (GWR) was developed to solve these problems (Fotheringham et al., 2002). It generates a spatially weighted model for each sample point or polygon centroid. This model usually yields a better goodness-of-fit than the global model. This concept can be generalized and applied to create geographically weighted statistics such as geographically weighted odds ratio, geographically weighted logistic regression, geographically weighted Poisson regression, etc. We applied geographically weighted logistic regression (GWLR) in this study.

2.2.1 Logistic Regression

Logistic regression is derived from the odds of a binary variable. Odds are the ratio of the probability that an event will happen to the probability that an event will not happen. A logistic regression model is

$$\log(odds) = logit\big(P(y=1)\big) = \ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

where y is a binary dependent variable, $x_1, x_2, \ldots, x_n$ are independent variables, which can be either binary or numerical, $b_1, b_2, \ldots, b_n$ are coefficients in the model (modified from Bewick et al., 2005).

A more common form of a logistic regression model is:

$$P(y=1) = \frac{1}{1+e^{-z}} \; or \; \frac{e^z}{1+e^z}$$

where

$$z = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

(modified from Bewick et al., 2005)

Predictions for the dependent variable can be made based on the value of P(y=1). When the value of P(y=1) is below 0.5, it is categorized as 0, whereas when the value of P(y=1) is above 0.5, it is categorized as 1.

When a dataset of k observations ($[(y_1, x_{11}, \dots, x_{1n}), (y_2, x_{21}, \dots, x_{2n}), \dots, (y_k, x_{k1}, \dots, x_{kn})]$) is available, to calibrate the coefficient for a logistic regression model, any coefficient $b_1, b_2, \dots, b_n$ can be obtained by solving:

$$b = (X^T X)^{-1} X^T Y, \text{ where where } X = \begin{bmatrix} x_{11} & \cdots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1k} & \cdots & x_{nk} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix}, b = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

(Fotheringham et al., 2002)

Logistic regression is able to not only incorporate multiple variables into a model that explains a binary dependent variable, but also estimate the odds ratio between the dependent variable and each independent variable while considering the effect from other variables. Thus, it is often employed to estimate adjusted odds ratios in epidemiological studies. For an independent variable x_i, the adjusted odds ratio can be estimated using $\exp(b_i)$:

$$\exp(b_i) = \frac{P(y = 1 | x_i = 1, x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n) / P(y = 0 | x_i = 0, x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n)}{P(y = 0 | x_i = 1, x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n) / P(y = 1 | x_i = 0, x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n)}$$

2.2.2 Geographically Weighted Logistic Regression

In the model above, only one globally applied model will be generated from the analysis of all study cases. However, in order to characterize the spatial non-stationarity of the association, multiple models can also be generated based on geographic locations. Thus, for a location i, the probability that the outcome of y is positive can be estimated based on a geographically weighted logistic regression (GWLR) model:

$$P(y_i) = \frac{1}{1 + e^{-z_i}} \; or \; \frac{e^{z_i}}{1 + e^{z_i}}$$

where

$$z_i = b_{i0} + b_{i1}x_1 + b_{i2}x_2 + \cdots + b_{in}x_n$$

(modified from Bewick et al., 2005)

For each location i, a unique logistic regression is generated. Its parameters are calibrated by nearby cases and weighted by a geographical weight function. Generally, cases closer to location i are given greater weights than those which are farther. For example, for location i with a dataset of k observations (as the example above), these parameters can be estimated by solving:

$b_i = (X^T W_i X)^{-1} X^T W_i Y$, where $W_i = \begin{bmatrix} w_{i1} \\ \vdots \\ w_{ik} \end{bmatrix}$

(Fotheringham et al., 2002)

$w_{i1}, w_{i2}, \ldots, w_{in}$ are the weights assign to cases 1 to k when calculating the coefficients of the GWLR model at location i. Within this model, the weight for each case is assigned based on the physical distance between location i and the location of each case. To transform a distance to a weight, several weight functions can be employed. In this study, we employed the bi-square kernel function, which is often applied in geographically weighted statistics (Fotheringham et al., 2002):

$$w_{ij} = \begin{cases} \left[1 - \left(\frac{d_{ij}}{h}\right)^2\right]^2 & if\ d_{ij} < h \\ 0 & otherwise \end{cases}$$

where $j = 1,2, \ldots, k$; d_{ij} is the distance between location i and observation j; h is bandwidth, which defines the maximum distance between location i and observation j to be considered in this model.

In this study, we generated a unique logistic regression model for each block group in Massachusetts. Thus, the i in the equation above refers to the centroid of a block group. Ideally 5,047 unique logistic regression models would be generated for the 5,047 block groups in the state.

The bandwidth in GWR or any geographically weighted regression can be calibrated by cross-validation (CV), generalized cross-validation criterion (GCV), Akaike Information Criterion (AIC),

Schwartz Information Criterion (SIC), etc. The goal of calibrating the bandwidth is to find the one bearing the best goodness-of-fit. We employed the approach of CV, which seeks a bandwidth that minimizes the quantity of

$$z = \sum_{i=1}^{k} [y_i - \hat{y}_i(b)]^2$$

where $\hat{y}_i(b)$ is the fitted value of y_i using a bandwidth of b (Fotheringham et al., 2002).

In our study, due to the limitation of computational capacity, we randomly selected 35 cases out of 584,603 and applied the bandwidths of 1,000, 2,000, 3,000,…, 10,000 meters. Among these ten bandwidths, we observed two minimums of z at 5,000 and 9,000 meters. Based on our visual observation on the shape of block groups, we can infer that this is due to the variation in block group sizes. Therefore, both 5,000 meters and 9,000 meters were applied. The former was applied to all block groups in the state. However, in rural areas, 5,000 meters might not be large enough to include sufficient cases in order to calculate coefficient in a logistic regression model. Thus, the later was employed only for larger block groups in the rural area. In this study, the rural area is composed of 533 block groups where the population density is less than 500 per square kilometer.

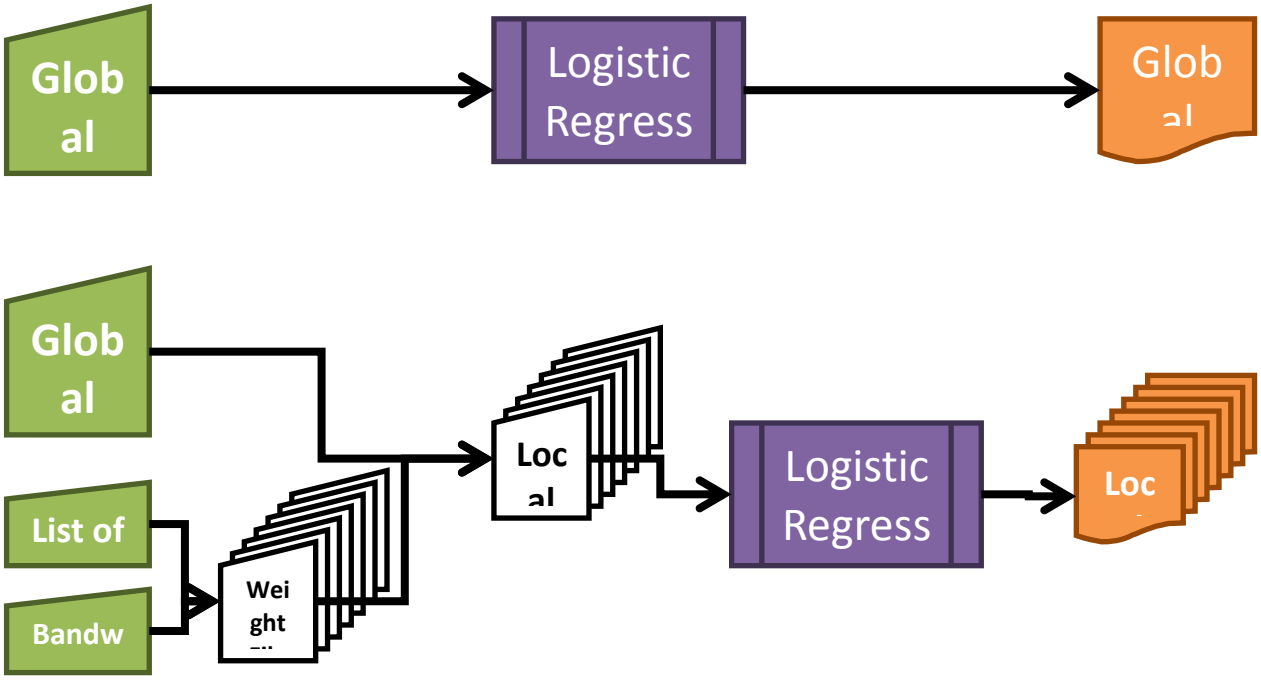
2.2.3 Process of Calculation

We used the binary logistic regression module in SPSS to do our logistic regression analysis. For each block group, we created a GWLR model to predict the likelihood of the LBW cases.

To calculate GWLR each model point (the location for which a local model is generated) has a unique input table and an output model (Figure 2). In this study, every block group in the state is a model point. While we generated a model for each block group, the location of each case was still determined by the block within which the case is located. The cases and weights for each model point are determined by the bandwidth and weight file. The process of table generation and regression analysis can be run repeatedly for all model points. In order to better control the calculation, which

involved the repeated process for all of the 5,047 block groups that cannot be properly handled in the existing GIS programs, we composed a Python script to generate a table that can be imported to SPSS for each block group. The entire process is composed of calculating weights, joining tables, sorting variables and importing to SPSS. By repeating this standard process, an output of logistic regression is generated for each block group. This regression model involves all the cases within that block group and its neighboring block groups. We wrote another Python script to parse these reports and create a new table that can be visualized in ArcGIS. Finally, we created coefficients maps to visually demonstrate the spatial non-stationarity of the association.

Figure 2 The Process of Global Logistic Regression (upper) and Geographically Weighted Logistic Regression (lower)



3 Results

In this chapter, we examined the spatial distributions of estimated odds ratios and significances of each variable one by one. The significances are evaluated with the significances of Wald, an index

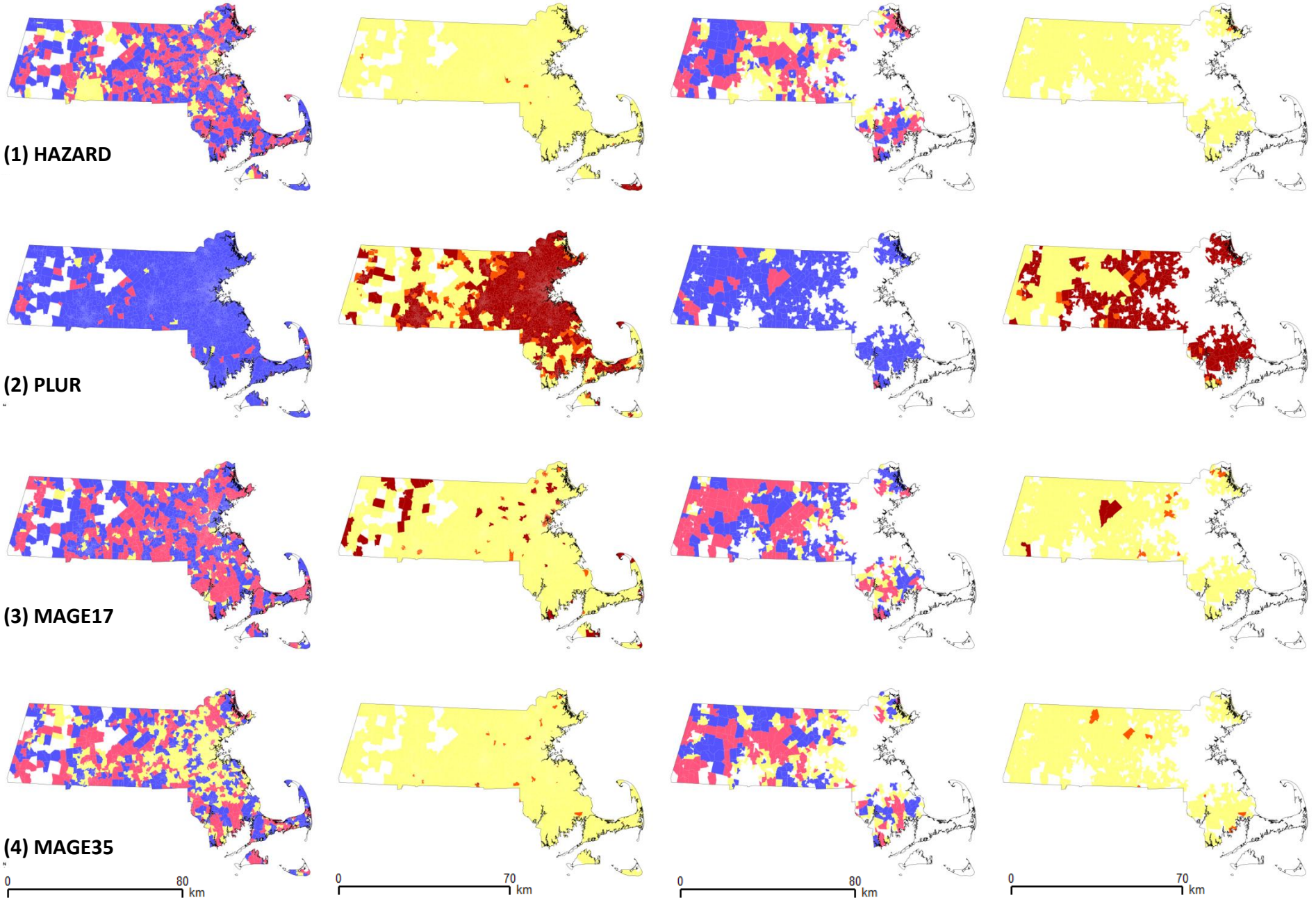

(1) HAZARD
(2) PLUR
(3) MAGE17
(4) MAGE35
0
80
km
0
70
km
0
80
km
0
70
km
(A) Adjusted Odds Ratio (bandwidth = 5,000m)
0.0 - 0.8
0.8 - 1.2
>1.2
(B) Significance of Wald (bandwidth = 5,000m)
<0.01
<0.05
0.05 - 1.00
(C) Adjusted Odds Ratio (bandwidth = 9,000m)
0.0 - 0.8
0.8 - 1.2
>1.2
(D) Significance of Wald (bandwidth = 9,000m)
<0.01
<0.05
0.05 - 1.00

Figure 3.1-4

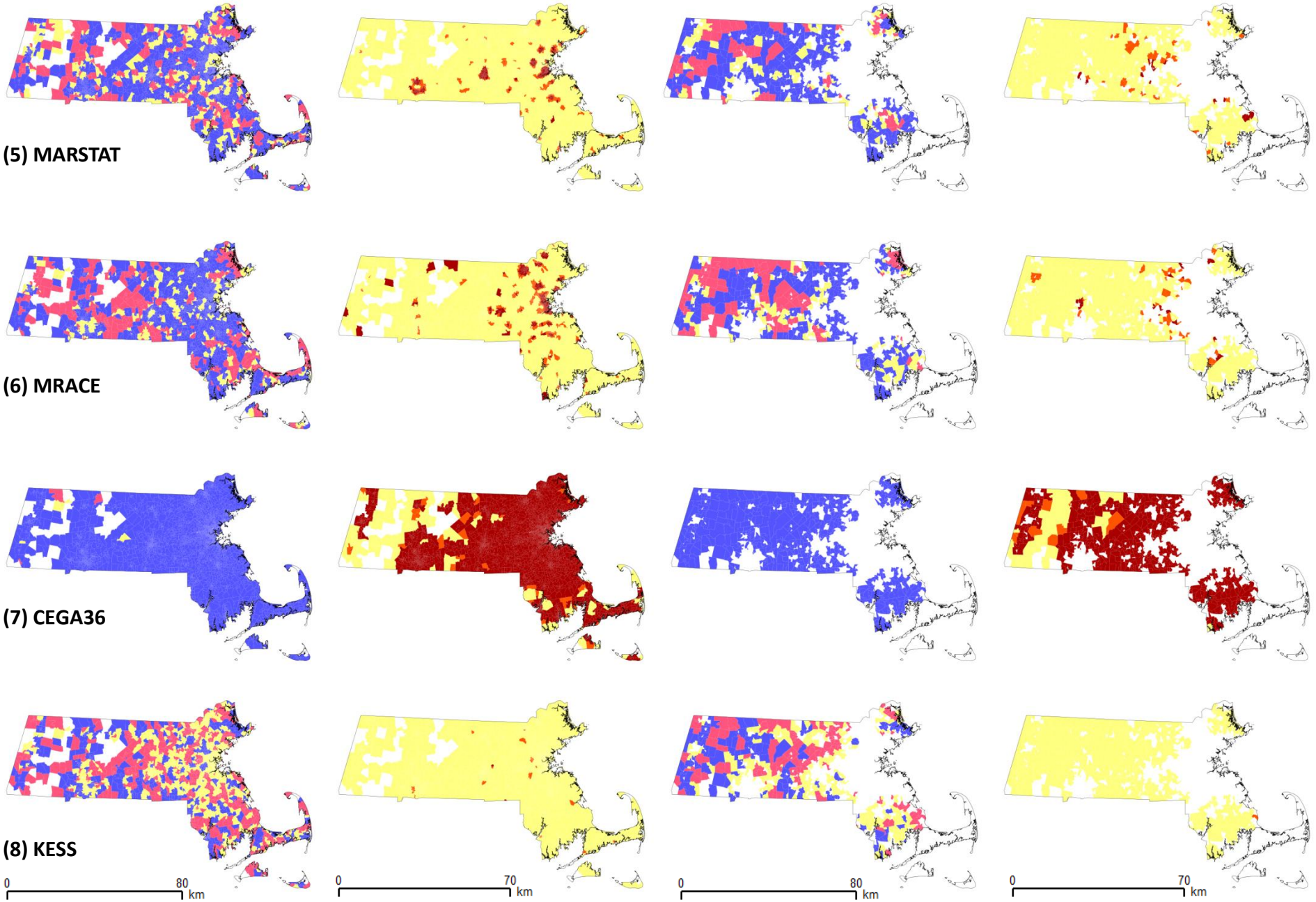

(5) MARSTAT
(6) MRACE
(7) CEGA36
(8) KESS
0
80
km
0
70
km
0
80
km
0
70
km
(A) Adjusted Odds Ratio (bandwidth = 5,000m)
0.0 - 0.8
0.8 - 1.2
>1.2
(B) Significance of Wald (bandwidth = 5,000m)
<0.01
<0.05
0.05 - 1.00
(C) Adjusted Odds Ratio (bandwidth = 9,000m)
0.0 - 0.8
0.8 - 1.2
>1.2
(D) Significance of Wald (bandwidth = 9,000m)
<0.01
<0.05
0.05 - 1.00

Figure 3.5-8

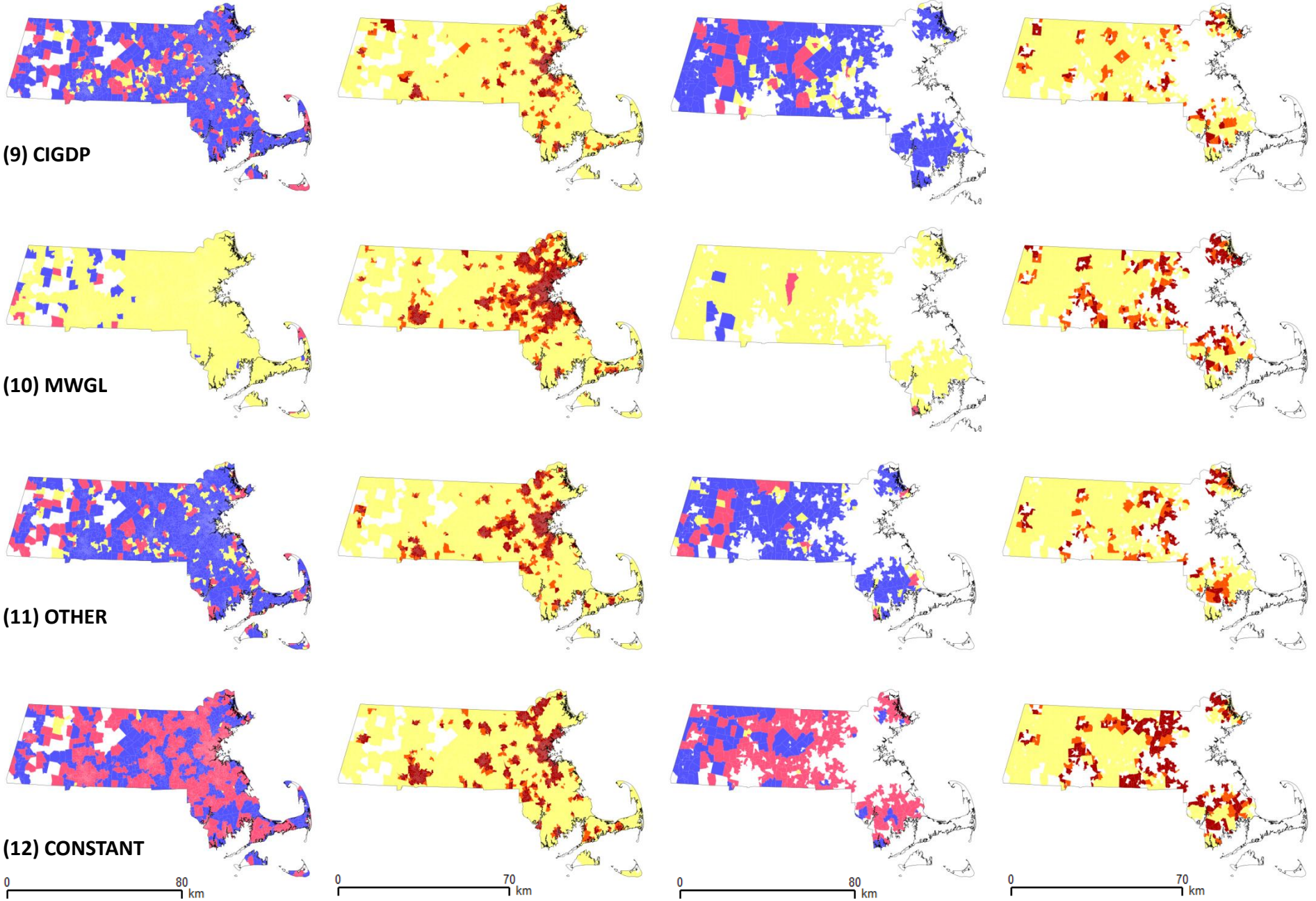

(9) CIGDP
(10) MWGL
(11) OTHER
(12) CONSTANT
0
80
km
0
70
km
0
80
km
0
70
km
(A) Adjusted Odds Ratio (bandwidth = 5,000m)
0.0 - 0.8
0.8 - 1.2
>1.2
(B) Significance of Wald (bandwidth = 5,000m)
<0.01
<0.05
0.05 - 1.00
(C) Adjusted Odds Ratio (bandwidth = 9,000m)
0.0 - 0.8
0.8 - 1.2
>1.2
(D) Significance of Wald (bandwidth = 9,000m)
<0.01
<0.05
0.05 - 1.00

Figure 3.9-12

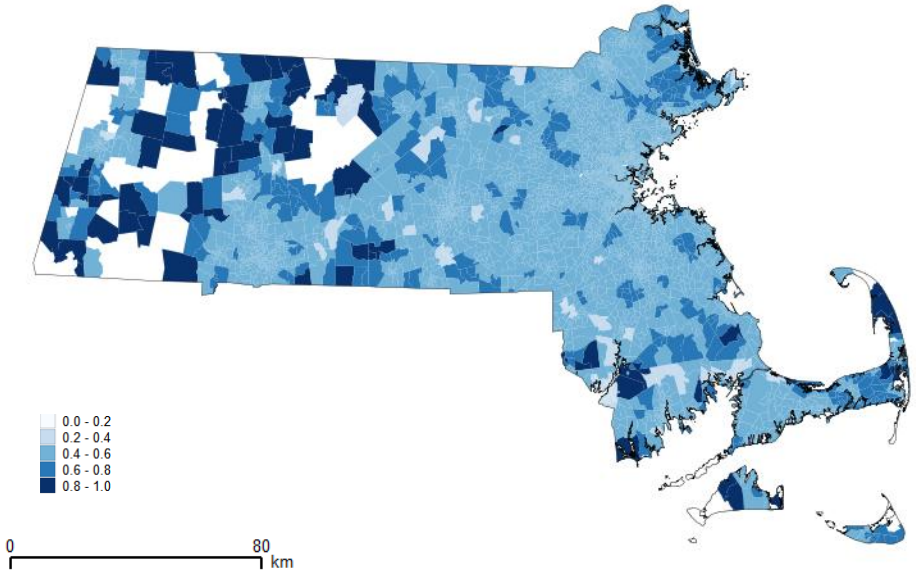


Figure 4a Nagelkarke R-Square (bandwidth = 5,000 m)

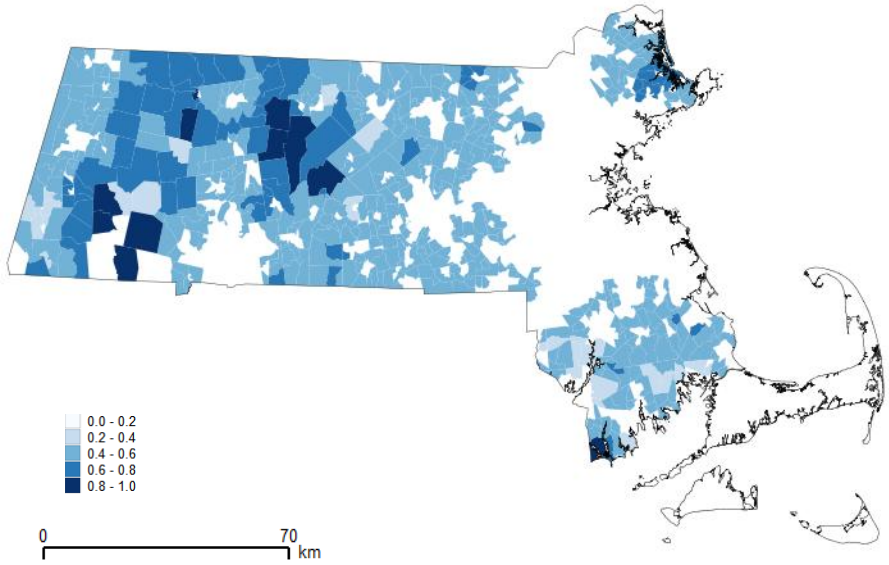


Figure 4b Nagelkarke R-Square (bandwidth = 5,000 m)

showing a variable's significance in a logistic regression model (Bewick et al., 2005). The maps of these results can be found in Figures 3 and 4.

Hazard-Related Score of Exposure to Lead (HAZARD) (Figure 3.1.a-d)

The distribution of adjusted odds ratio of HAZARD does not exhibit a regular pattern across the area of Massachusetts. But things change when we zoom in to the urban areas. No apparent association is found around Boston, Cambridge, northern Worcester and Springfield etc., while Watertown, Brookline, southern Worcester, etc. reported a higher risk of LBW with HAZARD. Local patterns also exist in other places, while the enumeration is omitted here. The significance of the adjusted odds ratio, nevertheless, indicates that it is only significant in 47 block groups at the level of 0.05 across the entire state.

Infant's Plurality (PLUR) (Figure 3.2.a-d)

PLUR is found to have an effect on the increased odd of LBW only in a few block groups across the entire state, these include areas around New Salem, Great Barrington, Hartsville, and a few block groups located in Cape Cod. The other areas are covered by adjusted odds ratios below 1.0, which indicate a negative association with LBW. The significance of this adjusted odds ratio indicates that the result is significant for 92.36% of the block groups at the bandwidth of 5,000 meters, and 83.52% of the block groups at the bandwidth of 9,000 meters. Those block groups are mostly located in central and eastern Massachusetts, with a few located in western Massachusetts around Sheffield, Pittsfield, Adams, Windsor, etc.

Mother's Age Below/Equal to 18 (MAGE17) (Figure 3.3.a-d)

Positive association between mother's age less than or equal to 17 and LBW is found to be in block groups irregularly distributed across Massachusetts, so is negative association. For some areas in central and western Massachusetts, mostly rural, there is no case with mother's age less than or equal

to 18, thus these areas have no results of adjusted odds ratio of MAGE17. In major cities such as Boston, Worcester, and Springfield, the association appears to be mostly negative. 11.7% of the results are significant at the level of 0.05 with a bandwidth of 5,000 meters, while 3.26% of the results are significant with the bandwidth of 9,000.

Mother's Age Above/Equal to 35 (MAGE35) (Figure 3.4.a-d)

Based on the value of adjusted odds ratio of MAGE35, positive association, as well as negative association, are found to be dispersed across the entire state. Several block groups in central Massachusetts, mostly in rural areas, reported a higher risk of LBW with MAGE35. The pattern in urban areas also differs. In Boston area, the association is either positive or not apparent. Only on the peripheral side, negative association exists. Unlike Boston, Springfield and Worcester has block groups with positive and negative association, as well as block groups that report no apparent association. The change is gradual and continuous. Only 63 in the entire state block groups get a significant value at the bandwidth of 5,000 meters, and 10 at the bandwidth of 9,000 meters.

Marital Status (MARSTAT) (Figure 3.5.a-d)

The results of adjusted odds ratio indicate that the associations between MARSTAT and LBW are mostly negative across the state of Massachusetts while block groups with positive association results scatter among them and away from urban centers. All three major urban areas, Boston, Worcester, and Springfield, appears to have negative association results, though the association in some small parts of Boston is not apparent. The results are mostly significant around urban areas such as Springfield, Worcester, Boston, Waltham, etc. In total, 29.06% of the block groups get a significant value at the bandwidth of 5,000 meters, and 9.8% of the block groups get a significant value at the bandwidth of 9,000 meters.

Mother's Race (MRACE) (Figure 3.6.a-d)

The distribution of adjusted odds ratio of MRACE is fairly patchy. In the extent of the entire state with the bandwidth 5,000 meters, no visible gradual change is found. Only in some specific areas the values are greater than one. These area scatter across the state and are generally away from the three major urban areas. With the bandwidth of 9,000 meters, we found a belt between Worcester and Springfield where there are higher odds of LBW if a mother's race is not white.

Gestational Age Equal to or Less than 36 Weeks (CEGA36) (Figure 3.7.a-d)

Aside from the odds ratios greater than 1.0 found in several block groups on the upper-left corner, most of the state is covered by low adjusted odds ratio values of CEGA36. In other words, early gestational age happened with a lower chance of LBW in most block groups. The bandwidth 5,000 and 9,000 meters show the similar patterns. The values of adjusted odds ratio vary across the state and show no gradual change.

Kess Index (KESS) (Figure 3.8.a-d)

KESS scores have a higher odds ratio with the LBW outcome at several areas away from urban centers, including Watertown and Molrose in the Boston urban area. These areas are scattered across the rural area of the state. The bandwidth 5,000 and 9,000 meters show the similar patterns. A gradual change cannot be found at the state scale.

Cigarettes during Pregnancy (CIGDP) (Figure 3.9.a-d)

Three areas of high adjusted odds ratio are found in north Worcester, north Springfield, and Newton Heights in south Boston. More block groups with high values are found in the rural area as well. The high and low values segregate in all three major urban areas. Two bandwidths results in the similar pattern. A gradual change cannot be found at the state scale. Using hazard and risk scores resulted in the similar pattern. A gradual change cannot be found at the state scale.

Mother's Weight Gain/Loss (MWGL) (Figure 3.10.a-d)

Based on the adjusted odds ratio of mother's weight gain or loss, odds ratios of positive association are found in Newton Heights, Fitchburg and north Worcester, Woburn, Lexington, etc. This variable was input as a numerical variable. The adjusted odds ratio indicates that LBW is more likely to associate with greater amount of weight gain. Most of the state does not show a positive association with LBW.

Other Health Conditions (OTHER) (Figure 3.11.a-d)

For this variable, high adjusted odds ratios are found in east Boston, south Worcester, Nowton and Natick, Cape Ann, etc. They either cover part of an urban area, or scatter in the rural area. For the rest of the state, this association is more likely to be negative.

4 Discussion

In the previous section, we examined the general spatial pattern of each variable. In this section, our focuses will be how well these variables explain LBW in different locations, how our study design affect the results we obtained, and how to improve the model we developed.

At first, in order to know how well these variables explain the LBW outcome, we considered the distribution of both the adjusted odds ratio and significance of each variable. We categorized variables according to their spatial patterns of adjusted odds ratio and significance. The model with the bandwidth 5,000 meters and the one with bandwidth 9,000 meters yield similar results. In terms of the distribution of adjusted odds ratio, we observed two types of spatial pattern: patchy and evenly negative. "Patchy" refers to the situation when the values larger than one (positive association) and those smaller than one (negative association) appear frequently while forming many clusters across the state. "Evenly negative" refers to the situation when the values smaller than one (negative association) appear much more frequently than those larger than one (positive association). In terms of the

distribution of significance of adjusted odds ratio, we observed three types of spatial pattern: low, partially high and high, which refer to the situation when significant values do not appear frequently, that significant values appear in urban areas, and that significant value appear across a large area in the state. Based on this typology, the eleven variables we incorporated plus the constant in the model can be divided as Table 5:

Table 5 The Typology of Spatial Patterns of Variables

		Overall Distribution of Significance		
		Low	Partially High	High
Overall Distribution of Adjusted Odds Ratio	Patchy	RISK, MAGE17, MAGE35, KESS	MARSTAT, MRACE, CIGDP, DISEASE, CONSTANT	
	Evenly negative		MWGL	PLUR, CEGA36

Although the variables MWGL, PLUR, and CEGA36 were proven to be associated with LBW based on the odds ratios and adjusted odds ratios calculated globally, in the output of GWLR, however, they show positive associations only at a few locations. The estimated odds ratios of MWGL are mostly close to one and significant in urban areas, while those of PLUR and CEGA36 are mostly negative and yield high levels of significance in most of the state. In other words, although they are apparently associated with LBW in the global model, they do not bear frequent positive associations in our geographically weighted models using either the bandwidth of 5,000 or 9,000 meters.

The other eight variables (RISK, MAGE17, MAGE35, KESS, MARSTAT, MRACE, CIGDP, DISEASE, CONSTANT) and constant show either positive or negative association at specific spots in the state. They form quite a few clusters of either positive or negative association across the state. Although not all of them yield a positive association at the same spot, the widths of these association clusters are fairly similar across the state, regardless of where they are and which variables they are based on. By measuring the widths of these clusters, we found that the widths are close to the

bandwidth we employed. Therefore, these patterns of cluster actually reflect the effect of bandwidth, as well as the importance of choosing an appropriate bandwidth for the weight function.

Four variables (RISK, MAGE17, MAGE35, KESS) yield strong associations at specific spots but their adjusted odds ratios are only significant in a few tiny areas. Although all of them might be effective in explaining LBW in some specific areas, these values are not quite reliable anywhere in the state.

Why do four of the eleven variables show patchy distributions and yield some high significance values around the urban area? According to our primary examination to the distribution of LBW (Figure 1), apparently there are several clusters of high LBW ratio in Massachusetts, including South Boston – Mattapan, downtown Worcester, downtown Springfield, downtown Pittsfield, Lowell, Ipswich, Brockton, etc. In these areas, the LBW ratios in population are extraordinarily high. Thus, when we tried to use the variables we incorporated to explain the occurrences of LBW in those areas, the common phenomena we know about those urban areas – high ethnic diversity, low percentage of majority population, inferior health status, etc – demonstrate their powers to explain the frequent occurrence of LBW, and yield the high values of significance as well. However, even if these variables are able to explain those frequent occurrences in urban areas, the constant in our model still demonstrate its strong association with LBW in these urban areas. This strong association implies that there are still more variables that can be incorporated into our model, particularly those which can reflect the social and demographic characteristics in urban environment.

Inferred from the model results, air pollution does not seem to be significantly correlated with LBW. Contrary to what we expected, high hazard scores in areas such as Springfield, Boston areas, and part of Worcester actually showed weak associations with LBW, (whereas areas indicating significant positive associations are small patches with low risk scores sporadically distributed across central and eastern Massachusetts). Different factors might have contributed to this pattern. First, the

risk score, according to EPA, was estimated with chemical toxicity at the TRI sites and the possible area of dispersion in the outdoor environment. It only measures the level of human outdoor exposure to air pollution. The effects from the indoor environment are not considered in RSEI. An exposure data that take human indoor and outdoor activities into account might demonstrate a stronger association with LBW. Second, the results might indicate that exposure to lead does not necessarily pose a negative effect on birth weight. Even though it has been indicated in several studies (such as Landgren, 1996), the association they indicated might not be constant everywhere.

To avoid producing a biased result, logistic regression has some requirements for sample size. It is recommended that with each increase in the number of independent variables, the number of cases in the rarer group should increase at least 10. In our study, this means that with every increase of independent variable, the number of LBW cases in the sample should increase at least 10. With 11 independent variables incorporated, the number of LBW cases should at least be equal to 110. This criterion was not satisfied in some blocks for the reason there is not as many as LBW cases within that area. Also, the total sample size should be above 500 in order to avoid overestimation of odds ratio. In our study, the sample size at some model points (block groups) failed to meet the criteria due to two major factors. One is edge effect. Since blocks at the edges of Massachusetts tend to have less neighboring cases, some of the blocks tend to have fewer cases than the required sample size. The other factor is bandwidth impact. This usually happens to large rural block groups with low population since the bandwidth is not large enough for the block group to include enough neighboring cases. The result is often including the cases just within the block group itself, which does not fulfill the requirement of an effective sample size. That is also one of the reasons why we get insignificant regression results in some block groups, and some even do not yield a result.

For major urban areas, the Nagelkerke R^2 varies around 45% - 60%, where we can consider that roughly 45% - 60% of the variance in the dependent variable is explained by the model (Figure 4).

High values mostly appear in large rural block groups in central and western Massachusetts, which indicates a fairly good fitness of the dataset and a corresponding higher overall correct classification percentage. This might be due to the reason that rural areas have fewer confounders compared with urban areas since the latter one is a much more complex social and ecological system. Factors such as mother's access to alcohol and drugs, mental stress, and various socioeconomic causes are not included in our analysis, and they all differ in urban and rural areas.

Another issue that cannot be ignored is the way we located each case. When precise residential addresses are available, researchers can utilize the US Census Topologically Integrated Geographic Encoding and Referencing (TIGER) database to locate each case by geocoding. In our dataset, instead, the finest geographic information of each case is census block. For each case, we assigned the centroid of the block where it is located as its geographic coordinates. These locations are not the actual residential locations of cases. This limitation of data brought two consequences in our study. First, the inaccurate geographic locations might impact the validation of our model. Depending on the sizes of blocks, the error of geographic locations can be from less than 10 meters to as much as 1,000 meters. Particularly greater errors can be expected in rural areas, where the sizes of blocks are relatively large. Second, assigning all cases in one block to the centroid of that block contributes to a high uncertainty in our model. The model is responsible for explaining various observations at one location. Consequently, using multiple observations at the same point to calibrate the coefficients in our model might have weakened the level of fitness.

5 Conclusions

In this study, we applied the concept of GWLR to build models which examined the association between LBW and selected independent variables over the entire Massachusetts. Different software such as ArcGIS, SPSS, and Python programming were used to produce the results. Based on our

analysis, we can conclude that the infant's plurality and gestational age equal to or below 36 weeks are highly associated with LBW, whereas mother's marital status, race, smoking history, disease conditions, and weight gain or lost are partially associated with LBW. Hazard score of air pollution of lead, mother's age and prenatal care proves to have a varied, unevenly distributed weak association with LBW across the entire Massachusetts. However, the results produced might be biased due to some inevitable deficiencies in data properties. To further improve the analysis, we could cross validate each GWLR model to get a better idea of how the model performs.

References

Aguilera, I., Guxens, M., Garcia-Esteban, R., Corbella, T., Nieuwenhuijsen, M. J., Foradada, C. M., & Sunyer, J. (2009). Association between GIS-Based Exposure to Urban Air Pollution during Pregnancy and Birth Weight in the INMA Sabadell Cohort. *Environ Health Perspect*, *117*(8). National Institute of Environmental Health Sciences.

Ash, M., & Fetter, T. R. (2004). Who Lives on the Wrong Side of the Environmental Tracks? Evidence from the EPA's Risk-Screening Environmental Indicators Model. *Social Science Quarterly*, *85*(2), 441-462. Blackwell Publishing.

Ashdown-Lambert, J. R. (2005). A review of low birth weight: predictors, precursors and morbidity outcomes. *The Journal of the Royal Society for the Promotion of Health*, *125*(2), 76-83.

Barker, D. J. P. (1995). Fetal origins of coronary heart disease. *Bmj*, *311*(6998), 171.

Basu, R., Woodruff, T. J., Parker, J. D., Saulnier, L., & Schoendorf, K. C. (n.d.). Comparing exposure metrics in the relationship between PM2.5 and birth weight in California. *J Expo Anal Environ Epidemiol*, *14*(5), 391-396.

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. *Critical Care*, *9*(1), 112-118.

Bove, F. J., Fulcomer, M. C., Klotz, J. B., Esmart, J., Dufficy, E. M., & Savrin, J. E. (1995). Public Drinking Water Contamination and Birth Outcomes . *American Journal of Epidemiology* , *141* (9), 850-862.

Brauer, M., Lencar, C., Tamburic, L., Koehoorn, M., Demers, P., & Karr, C. (2008). A Cohort Study of Traffic-Related Air Pollution Impacts on Birth Outcomes. *Environ Health Perspect*, *116*(5). National Institute of Environmental Health Sciences.

Chan, T.-C., Chen, M.-L., Lin, I.-F., Lee, C.-H., Chiang, P.-H., Wang, D.-W., & Chuang, J.-H. (2009). Spatiotemporal analysis of air pollution and asthma patient visits in Taipei, Taiwan. *International Journal of Health Geographics*, *8*(1), 26.

Clougherty, J. E., Levy, J. I., Kubzansky, L. D., Ryan, P. B., Suglia, S. F., Canner, M. J., & Wright, R. J. (2007). Synergistic Effects of Traffic-Related Air Pollution and Exposure to Violence on Urban Asthma Etiology. *Environ Health Perspect*, *115*(8). National Institute of Environmental Health Sciences.

Cramer, J. (1995). Racial and Ethnic Differences in Birthweight: The Role of Income and Financial Assistance. *Demography*, *32*(2), 231-247.

Dugandzic, R., Dodds, L., Stieb, D., & Smith-Doiron, M. (2006). The association between low level exposures to ambient air pollution and term low birth weight: a retrospective cohort study. *Environmental Health: A Global Access Science Source*, *5*(1), 3.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships*.

Gouveia, N., Bremner, S. A., & Novaes, H. M. D. (2004). Association between ambient air pollution and birth weight in São Paulo, Brazil . *Journal of Epidemiology and Community Health , 58* (1), 11-17.

Gryparis, A., Coull, B. A., Schwartz, J., & Suh, H. H. (2007). Semiparametric latent variable regression models for spatiotemporal modelling of mobile source particles in the greater Boston area. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 56*(2), 183-209.

Ha, E.-H., Hong, Y.-C., Lee, B.-E., Woo, B.-H., Schwartz, J., & Christiani, D. C. (2001). Is Air Pollution a Risk Factor for Low Birth Weight in Seoul? *Epidemiology, 12*(6).

Hart, J. E., Yanosky, J. D., Puett, R. C., Ryan, L., Dockery, D. W., Smith, T. J., Garshick, E., et al. (2009). Spatial Modeling of PM_{10} and NO_2 in the Continental United States, 1985–2000. *Environ Health Perspect, 117*(11).

Hattersley, A. T., & Tooke, J. E. (1999). The fetal insulin hypothesis: an alternative explanation of the association of low bir thweight with diabetes and vascular disease. *The Lancet, 353*(9166), 1789-1792.

Jerrett, M., Burnett, R. T., Ma, R., Pope, C. A. I. I. I., Krewski, D., Newbold, K. B., Thurston, G., et al. (2005). Spatial Analysis of Air Pollution and Mortality in Los Angeles. *Epidemiology, 16*(6).

Kannan, S., Misra, D. P., Dvonch, J. T., & Krishnakumar, A. (2007). Exposures to airborne particulate matter and adverse perinatal outcomes: a biologically plausible mechanistic framework for exploring potential. *Ciência & Saúde Coletiva, 12*(6), 1591-1602.

Krieger, N. (1992). Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology. *American Journal of Public Health, 82*(5), 703-710.

Landgren, O. (1996). Environmental pollution and delivery outcome in southern Sweden: a study with central registries. *Acta Paediatrica, 85*(11), 1361-1364.

Lee, B. E., Ha, E. H., Park, H. S., Kim, Y. J., Hong, Y. C., Kim, H., & Lee, J. T. (2003). Exposure to air pollution during different gestational phases contributes to risks of low birth weight . *Human Reproduction , 18* (3), 638-643.

Liao, D., Peuquet, D. J., Duan, Y., Whitsel, E. A., Dou, J., Smith, R. L., Lin, H.-M., et al. (2006). GIS Approaches for the Estimation of Residential-Level Ambient PM Concentrations. *Environ Health Perspect, 114*(9).

McEntee, J. C., & Ogneva-Himmelberger, Y. (2008). Diesel particulate matter, lung cancer, and asthma incidences along major traffic corridors in MA, USA: A GIS analysis. *Health & Place, 14*(4), 817-828.

Mosby. (2009). *Mosby's medical dictionary, 8th edition*. Mosby.

Nethery, E., Brauer, M., & Janssen, P. (2008). Time–activity patterns of pregnant women and changes during the course of pregnancy. *Journal of Exposure Science and Environmental Epidemiology*, *19*(3), 317-324.

Ozdenerol, E., Williams, B., Kang, S. Y., & Magsumbol, M. (2005). Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters. *International Journal of Health Geographics*, *4*(1), 19.

Pattenden, S., Dolk, H., & Vrijheid, M. (1999). Inequalities in low birth weight: parental social class, area deprivation, and "lone mother" status. . *Journal of Epidemiology and Community Health , 53* (6), 355-358.

Ross, Z., English, P. B., Scalf, R., Gunier, R., Smorodinsky, S., Wall, S., & Jerrett, M. (2005). Nitrogen dioxide prediction in Southern California using land use regression modeling: potential for environmental health analyses. *J Expos Sci Environ Epidemiol*, *16*(2), 106-114.

Ryan, P H, LeMasters, G. K., Biswas, P., Levin, L., Hu, S., Lindsey, M., Bernstein, D. I., et al. (2007). A comparison of proximity and land use regression traffic exposure models and wheezing in infants. *Environmental health perspectives*, *115*(2), 278.

Ryan, Patrick H, & LeMasters, G. K. (2007). A Review of Land-use Regression Models for Characterizing Intraurban Air Pollution Exposure. *Inhalation Toxicology*, *19*(s1), 127-133.

Shah, P. S., & Balkhair, T. (2011). Air pollution and birth outcomes: A systematic review. *Environment International*, *37*(2), 498-516.

Slama, R., Morgenstern, V., Cyrys, J., Zutavern, A., Herbarth, O., Wichmann, H.-E., & Heinrich, J. (2007). Traffic-Related Atmospheric Pollutants Levels during Pregnancy and Offspring's Term Birth Weight: A Study Relying on a Land-Use Regression Exposure Model. *Environ Health Perspect*, *115*(9).

Southern California Center for Reproductive Medicine. (2011). *A Woman's Age and Fertility*, Southern California Center for Reproductive Medicine. http://www.socalfertility.com/age-and-fertility.html . Retrieved on 1/2/2011.

US Environmental Protection Agency. (2011). National Ambient Air Quality Standards (NAAQS). *National Ambient Air Quality Standards (NAAQS)*.

Villanueva, C. M., Durand, G., Coutté, M.-B., Chevrier, C., & Cordier, S. (2005). Atrazine in municipal drinking water and risk of low birth weight, preterm delivery, and small-for-gestational-age status . *Occupational and Environmental Medicine, 62* (6), 400-405.

Wilhelm, M., & Ritz, B. (2005). Local Variations in CO and Particulate Air Pollution and Adverse Birth Outcomes in Los Angeles County, California, USA. *Environ Health Perspect*, *113*(9). National Institute of Environmental Health Sciences.

Witkowski, K. M., & Johnson, N. E. (1992). Organic-solvent water pollution and low birth weight in Michigan. *Biodemography and Social Biology*, *39*(1-2), 45-54.

World Health Organization. (1976). WHO: recommended definitions, terminology and format for statistical tables related to the perinatal period and use of a new certificate for cause of perinatal deaths. *Acta Obstet Gynecol Scand 56*: 247–253.

World Health Organization. (2010). Nutrition Landscape Information System (NLIS) country profile indicators: interpretation guide. *Geneva: World Health Organization*.

Yanosky, J. D., Paciorek, C. J., Schwartz, J., Laden, F., Puett, R., & Suh, H. H. (2008). Spatio-temporal modeling of chronic PM10 exposure for the Nurses' Health Study. *Atmospheric Environment*, *42*(18), 4047-4062.

Yu, H.-L., Chen, J.-C., Christakos, G., & Jerrett, M. (2008). BME Estimation of Residential Exposure to Ambient PM_{10} and Ozone at Multiple Time Scales. *Environ Health Perspect*, *117*(4). National Institute of Environmental Health Sciences.