Economics and Human Biology 10 (2012) 43–53

Contents lists available at SciVerse ScienceDirect

Economics and Human Biology

journal homepage: http://www.elsevier.com/locate/ehb

Height and death in the Antebellum United States: A view through the lens of geographically weighted regression

Dongwoo Yoo *

West Virginia University, Department of Economics, 1601 University Ave., PO Box 6025, Morgantown, WV 26506-6025, USA

ARTICLE INFO

Article history:
Received 30 September 2011
Accepted 30 September 2011
Available online 11 October 2011

Keywords:
Geographically weighted regression
Spatial patterns
Antebellum puzzle
Physical stature
Height
Anthropometric history United States

ABSTRACT

Factors related to geography such as climate, natural resources or waterways often affect human activities. However, traditional approaches such as ordinary least squares (OLS) have limitations in investigating such patterns. Unlike OLS regression, geographically weighted regression (GWR) allows the coefficients of explanatory variables to differ by locality by giving relatively more weight to geographically close observations. GWR depicts spatial patterns. This paper examines the pattern of height and crude death rate in the United States prior to the Civil War by this method. The GWR results show that access to water transportation increased mortality and decreased stature in the food exporting areas of the Midwest, and the opposite pattern appeared in the food importing areas of the Northeast.

1. Introduction

Factors related to geography such as climate, natural resources or waterways often affect human activities. Inevitably, some regions were more favorably situated than others, and economic development has geographic components. In the U.S., for example, industrialization proceeded unevenly across the country, as some regions led the process and others lagged (Easterlin, 1961; Cuff, 2005).[1] Researchers frequently use geographic variables in order to analyze observable outcomes. For example, regional dummies, distance, and latitude are used to estimate the effect of geography. However, traditional approaches such as ordinary least squares (OLS) regression have limitations in investigating geographical patterns. OLS assumes that the coefficients of the independent variables are constant within a region, thereby omitting fine-grained spatial information of observations by estimating an average effect. To be sure, the average value does provide a meaningful summary if there is little variation within the defined space. However, given spatially differentiated economic activities, the global statistic may not accurately reflect local conditions. Regional dummy variables still treat regions as homogenous and other measures of geography such as distance and latitude also impose a type of homogeneity. In addition, the definition of regions can also influence the outcome of the analysis. In other words, the methods used so far in order to understand spatial variation should be improved upon.

Geographically weighted regression (GWR) is a statistical method well-adapted to study fine-grained spatial patterns. The method gives relatively more weight to close observations and diminishing weight to distant ones. In other words, GWR uses primarily geographically close

* Tel.: +1 304 293 0968; fax: +1 304 293 5652.
E-mail address: dongwoo.yoo@mail.wvu.edu.

[1] For example, short term interest rates varied widely in the postbellum years and converged slowly (Bodenhorn and Rockoff, 1992). In the nineteenth century, fertility differed across the regions, whereby child–woman ratios tended to increase from east to west (Carter et al., 2004). The public funding for higher education and access to public colleges and universities varied significantly among states from the late 1890s to the 1940s (Goldin and Katz, 1999). The diffusion of hybrid corn differed across regions in the 1930s. Some regions planted hybrids earlier than others, and some regions, once the shift began, made the transition more rapidly than others (Griliches, 1960). In many democratic countries, voters have regional political attitudes.

1570-677X/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ehb.2011.09.006

observations to estimate local coefficients. This spatial weighting scheme is based on the notion that using geographically close observations is the best way to estimate the local coefficients. Unlike OLS, the technique allows the coefficients of explanatory variables to vary across space, that is to say, to differ by locality, in our case the county. Moreover, it requires few prior assumptions on the division of regions and has advantages in capturing the non-linear geographic effects. Consequently, GWR is used in analyzing spatial patterns and making and testing spatial hypotheses in many fields, though not yet in economics (Fotheringham et al., 2002; Wang et al., 2008).

Here I illustrate the method using height and the crude mortality rate in the Antebellum United States as a case study. The U.S. economy grew robustly from 1830 to 1860 (0.9–1.3% per annum), but other indicators of the quality of life, such as poverty rates and especially mortality rates and nutrition, experienced major setbacks. Average height decreased from 1830 to 1860 by about 3 cm (from 173.5 cm to 170.6 cm) or 1.2 in. (Haines et al., 2003; Zehetmayer, 2011; A'Hearn, 1999). The decline in adult height that accompanied the onset of modern economic growth prior to the outbreak of the Civil War is referred to as the Antebellum Puzzle insofar as height and income were expected to move in unison rather than counter cyclically.

Anthropometric research provides a framework for analyzing a population's history of net nutrition during childhood and adolescence, or diet minus claims on the diet made by basal metabolism, work, and disease. Likely economic changes that would affect nutritional status in this period include urbanization, population growth, industrialization, commercialization, higher food prices, the rise of public schools that facilitated the spread of diseases among children and at the same time decreased their work effort, the transportation revolution that spread diseases and lowered the prices of manufactured goods purchased by farm families and raised the prices they received for food sold off the farm; and the Civil War, which disrupted food production and also spread diseases.[2] Of course, several explanations are interrelated, which makes it difficult to measure the pure effect of a single variable. Lower transportation costs, for example, influenced urbanization, relative prices of food and manufactured goods and the spread of pathogens through migration.

Komlos suggested that the deterioration in nutritional status was likely caused by forces endogenous to economic development such as the increased relative and absolute food prices which induced a substitution away from meat and dairy product consumption (1987, p. 907). In other words, exogenous causes such as cholera epidemics were not likely to have caused such widespread decline in physical stature because the epidemics were episodic and would not have given rise to such long-term trends. "Besides, when it did strike, cholera affected relatively few individuals, and because of the high mortality rate, not many of those who contracted the disease would have enrolled in West Point". Subsequently, Craig and Weiss (1998) and Haines et al. (2003) regressed height on food output, transportation, wealth, rate of urbanization, and rate of foreign born at the county level and concluded that the decline in height was caused by the development of transportation, urbanization, and also a deteriorating diet thereby confirming Komlos' hypothesis although they did not rule out complementary effects from the disease environment (Komlos and Coclanis, 1997; Cuff, 2005). In particular, they found that physical stature at the local level correlated positively with local protein and calorie production after controlling for wealth, occupation, transportation access and the disease environment. This implies that local prices mattered to food consumption.[3]

Previous work on this topic used OLS estimates, which assumes that the effect various independent variables (such as distance to water transportation) on mortality and physical stature were homogeneous throughout the country or at least within large regions of it. This assumption is questionable for the Antebellum era as the implication of economic processes varied considerably over the country at the onset of modern economic growth. For instance, water transportation was used for exporting food products in the Midwest, and for importing food in the Northeast. Therefore, the implication of water transportation on health was probably completely different in the two regions.

The coefficients of an OLS regression, however, are unable to capture this complex spatial pattern as it would reflect only the average of the two effects unless one used dummy variables for the regional effect with interaction terms for distance to water transportation. But even such a strategy would miss the intra-regional dispersion of the strength of the relationship between height (mortality) and distance from water transportation insofar as that relationship would be expected to vary even within the larger regions themselves. In contrast, GWR can estimate complicated spatial patterns, and is thus able to capture the differential regional effects of transportation on nutritional status. It accomplishes this more elegantly than would a dummy variable approach.

2. Geographically weighted regression

Suppose that one wants to estimate the coefficient at location 1 in Fig. 1. OLS gives the same weight to all locations and estimates a global estimate which is identical across all locations. In contrast, GWR gives the greatest weight to location 1, large weight to locations 2 and 4 (close observations), small weight to locations 3, 5, and 7 (more distant observations), and ignores locations 6, 8, and 9 (very distant observations). Thus, GWR estimates a specific coefficient for location 1, that is to say, every location obtains its own coefficient.[4] In other words, the

[2] For a summary see Steckel (1995) and Komlos (1987, 1998).

[3] At the same time the height of slaves was not contingent on local food prices because "slave owners had a strong interest in monitoring and controlling the diet of their slaves" (Haines et al., 2011; Komlos and Coclanis, 1997).

[4] The spatial autoregressive (SAR) model is different from GWR because it estimates global rather than local coefficients. According to Wang et al., "it may be concluded that GWR is being established as a standard tool in exploratory spatial-data analysis due to its effectiveness and wide applications" (Wang et al., 2008).

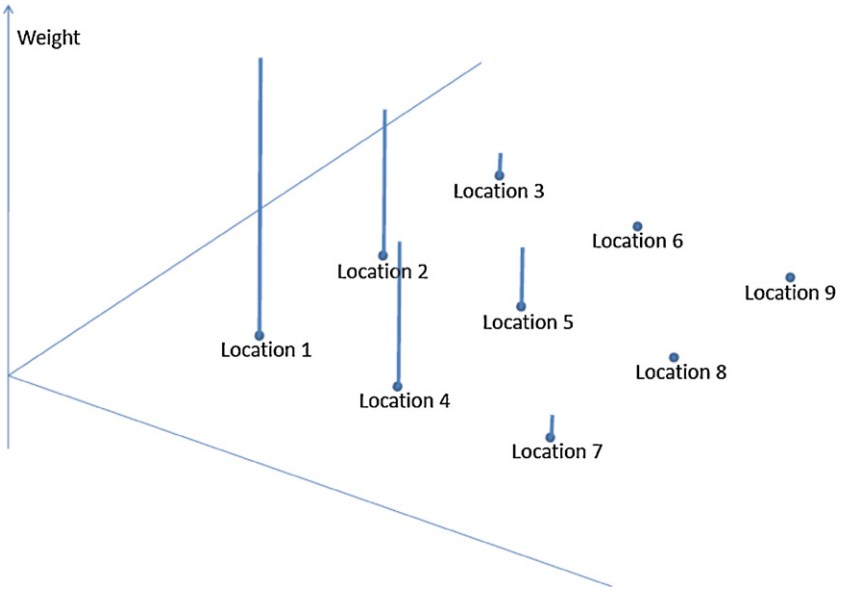

Fig. 1. Location index.

estimated coefficient of transportation, say, at Chicago may differ from that at New York not only because the average impact of geography could be different across the regions but also because the nearby observations and weights are different.

In general, a GWR estimator of $b(m) = (b_1(m),\ldots,b_r(m))$ can be written as follows:

$$\hat{b}(m) = \arg\min_{b(m)} \sum_{k=1}^{n} w_k(m)(y_k - x_k' b(m))^2$$

where 'arg min' denotes an argument that minimizes the function it precedes, y_k is a dependent variable(in our case height or mortality), and x_k is an r dimensional column vector of independent variables, $b(m)$ is r dimensional column vector (GWR estimator), n is the number of total observations, m is the location of interest, and $w_k(m)$ is a weighting function at location m ($m = 1,\ldots,1251$ and $r = 7$ in the crude death rate regression; $m = 1,\ldots,6557$ and $r = 10$ in the height regression).

A quadratic kernel is an example of a non-parametric weighting function widely used in GWR:

$$w_k(m) = \left(1 - \left(\frac{\rho(m,k)}{h}\right)^2\right)^2, \quad \text{if } \rho(m,k) < h$$
$$= 0, \quad \text{otherwise}$$

where $\rho(m, k)$ is the distance between the location of interest m and the location k, and h is the bandwidth, which is the distance between the point of interest and the closest observation given zero weight. In the previous example at location 1, if the bandwidth is $\sqrt{5}$ (assuming that the distance between points 1 and 2 is equal to 1 unit, then the distance between points 1 and 6 is $\sqrt{5}$), then the quadratic kernel assigns weight of 1 to location 1, 16/25 to locations 2 and 4, 9/25 to location 5, and 1/25 to locations 3 and 7.

GWR is relatively insensitive to the choice of weighting function if the weight decreases as the distance between two points increases (Fotheringham et al., 2002), but is sensitive to the bandwidth. There are two ways to determine the bandwidth of GWR. A fixed kernel uses a given bandwidth, which does not vary with data density. Thus, the number of observations used in estimation differs according to data density. In other words, more observations are used in the area where observations are dense than the area where they are sparse. An adaptive kernel uses a fixed number of observations. Thus, the bandwidth differs according to data density, and larger bandwidth is used in the area where observations are sparse than in the area where observations are dense. In general, as the number of observations increases the bandwidth decreases because the observations are more geographically compact. If the data density is homogenous across space, the fixed kernel – capturing regional characteristics in a consistent way – is recommended. However, if the data density varies significantly across the space, then the adaptive kernel – adjusting the number of observations according to data density – is recommended.

Empirically, the optimal bandwidth is determined by minimizing cross validation, $\mathrm{CV} = \sum_{i=1}^{n} (y_i - \hat{y}_{\neq i})^2$ where $\hat{y}_{\neq i}$ is the fitted value of y_i with the observations for point i omitted from the calibration process.

Statistical properties such as unbiasedness and consistency of GWR are similar to non-parametric estimation techniques, especially locally weighted regressions (Cleveland, 1979). Compared to the random coefficient model, GWR assumes that the coefficient $b(m)$ is a (continuous) function of location. However, because the spatial distribution of $b(m)$ is not known, non-parametric methods are used for the estimation.

A user-friendly software for GWR is available from National Centre for Geocomputation National University of Ireland (stewart.fotheringham@may.ie). Recently, the

Table 1
OLS regression results (crude death rates).

Variable	(1)	(2)	(3)	(4)	(5)
Intercept	13.44*** (0.50)	5.74*** (0.86)	4.10*** (1.02)	11.31*** (1.74)	13.79*** (1.89)
URBAN	13.16*** (2.35)	13.30*** (2.25)	13.96*** (2.30)	11.68*** (2.41)	15.51*** (3.33)
FOREIGN	1.20 (2.92)	14.72*** (3.02)	15.59*** (3.28)	16.98*** (3.32)	19.30*** (2.27)
MALARIA	–	21.16*** (1.95)	17.69*** (2.27)	19.19*** (2.28)	19.33*** (2.28)
NORTHEAST	–	–	REF	REF	REF
MIDWEST	–	–	2.14*** (0.83)	−4.94*** (1.82)	−5.63** (2.17)
SOUTH	–	–	2.69*** (0.89)	−6.04*** (1.82)	−9.57*** (2.06)
TRANSPORT	3.80*** (0.50)	4.06*** (0.48)	4.14*** (0.48)	−2.61 (1.50)	−1.07 (1.57)
TRANSPORT × MIDWEST	–	–	–	5.41*** (1.69)	4.03** (1.75)
TRANSPORT × SOUTH	–	–	–	8.48*** (1.60)	6.57*** (1.69)
CALORIES	0.10 (0.07)	0.14** (0.06)	0.10 (0.06)	−0.85*** (0.28)	−0.56* (0.29)
CALORIES × MIDWEST	–	–	–	0.91*** (0.30)	0.69** (0.32)
CALORIES × SOUTH	–	–	–	1.00*** (0.28)	0.63*** (0.30)
WEALTH	1.30*** (0.29)	2.09*** (0.29)	2.39*** (0.30)	2.52*** (0.30)	0.71 (0.68)
WEALTH × MIDWEST	–	–	–	–	0.38 (0.90)
WEALTH × SOUTH	–	–	–	–	2.91*** (0.78)
Adjusted *R*-squared	0.12	0.19	0.20	0.22	0.23
F	34.49***	51.11***	39.73***	30.17***	27.66***
N	1260	1251	1251	1251	1251

* Significant at 10% level.
** Significant at 5% level.
*** Significant at 1% level.

GWR Tool has been developed by ESRI (the Environmental Systems Research Institute of Redlands, CA) in ArcGIS 9.3. Also, code for GWR in R is available[5] from Chris Brunsdon (cb179@le.ac.uk) and some user-written commands in Stata are available[6] (http://www.staff.ncl.ac.uk/m.s.pearce/stbgwr.htm).

3. Empirical analysis

I first reproduce Haines et al.'s (2003) OLS regressions with county crude death rate as the dependent variable and report the coefficients (Table 1 column 1). In order to control the effect of disease, I add MALARIA to the Haines et al.'s (2003) specification (Table 1 column 2).[7]

Then I map the mean of the residuals at the county level (1251 observations) (Fig. 2). This enables us to observe the usefulness of the GWR model. Residuals are obviously not distributed randomly over space as the OLS model implicitly assumes. Large OLS residuals of the regression are clustered geographically, implying that the effect of the explanatory variables is not constant throughout the U.S. and therefore factors associated with geography were an important determinant of crude death rates. The spatial distribution indicates that there are large residuals along the Ohio and Mississippi River valleys and on the Atlantic Seaboard and negative residuals in the interior of the country. This is an indication that diseases were transmitted easier along waterways.

In order to estimate local effects, I expand Haines et al.'s (2003) OLS model on crude death rates[8] to GWR where all of the coefficients in the equation are indexed by m ($m = 1,\ldots,1251$; 1251 is the number of observations(the number of counties in the sample). GWR is equivalent to estimating 1251 different local regressions, thus it is different from fixed effects regression. The specification is similar to Haines et al. (2003), except that we allow the coefficients (b) of the independent variables to vary spatially:

$$\begin{aligned}\mathrm{CDR}_m = {} & b_1(m) + b_2(m)\ \mathrm{CALORIES}_m \\ & + b_3(m)\ \mathrm{WEALTH}_m + b_4(m)\ \mathrm{TRANSPORT}_m \\ & + b_5(m)\ \mathrm{URBAN}_m + b_6(m)\ \mathrm{FOREIGN}_m \\ & + b_7(m)\ \mathrm{MALARIA}_m + e_m \end{aligned} \quad (1)$$

where CDR_m is the crude death rate in the mth county in 1850, WEALTH_m is the sum of agricultural and industrial wealth (per capita, unit $100) in the county in 1850. $\mathrm{CALORIES}_m$ is the marketable surplus per person of calories in the county in 1840. $\mathrm{TRANSPORT}_m$ equals one if the county was on a navigable waterway in 1840, zero otherwise. URBAN_m is the proportion of the county's population living in an urban area in 1840. $\mathrm{FOREIGN}_m$ is the proportion of the county population in 1850 which was foreign born (Haines et al., 2003). $\mathrm{MALARIA}_m$ is the

[5] However, this code is not an exact replica of the Windows-based program.
[6] In Stata, type, "finditgwr" to locate and install the GWR command, which fits geographically weighted regression models.
[7] Controlling MALARIA increases *R*-squared from 0.12 (the original model; Table 1 column 1) to 0.19 (Table 1 column 2). This adjustment in specification indicates that FOREIGN has a downward omitted variable bias in the original model, which can be explained by the negative correlation between the proportion of foreign born and the estimated risk of malaria. At that time foreigners seldom settled in the South where the risk of malaria was high, but they preferred the North where the risk of malaria was low (the negative correlation between FOREIGN and MALARIA is expected). The significant coefficient on FOREIGN suggests that immigrants brought diseases to the county where they settled. Because the risk of malaria was unavailable in 9 counties, 1251 observations are used in the estimation.
[8] My analysis is based on the food output and Union Army data kindly provided by Michael Haines.

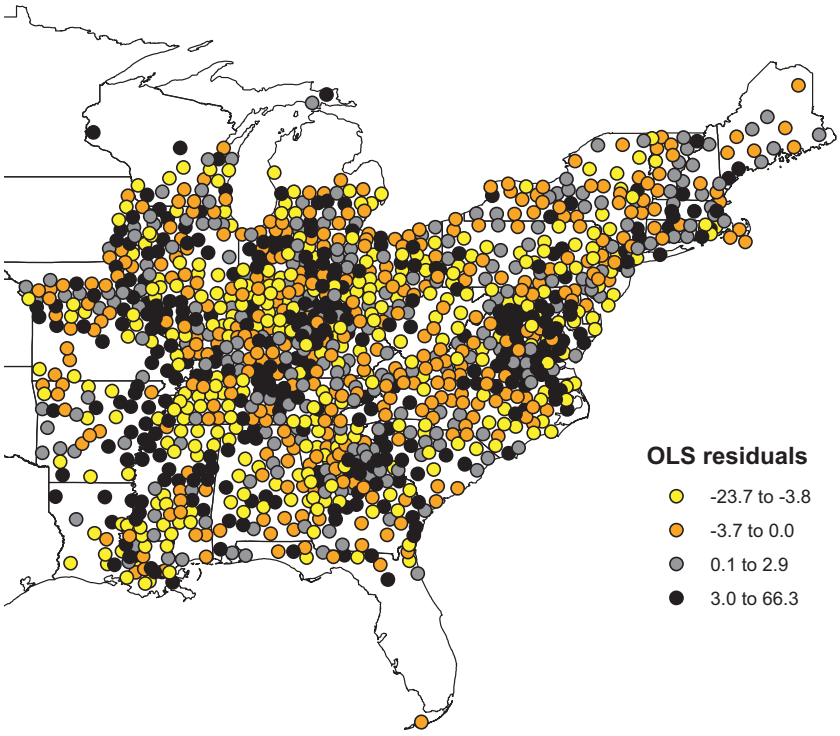

Fig. 2. Residuals of baseline OLS regression with CDR as dependent variable.

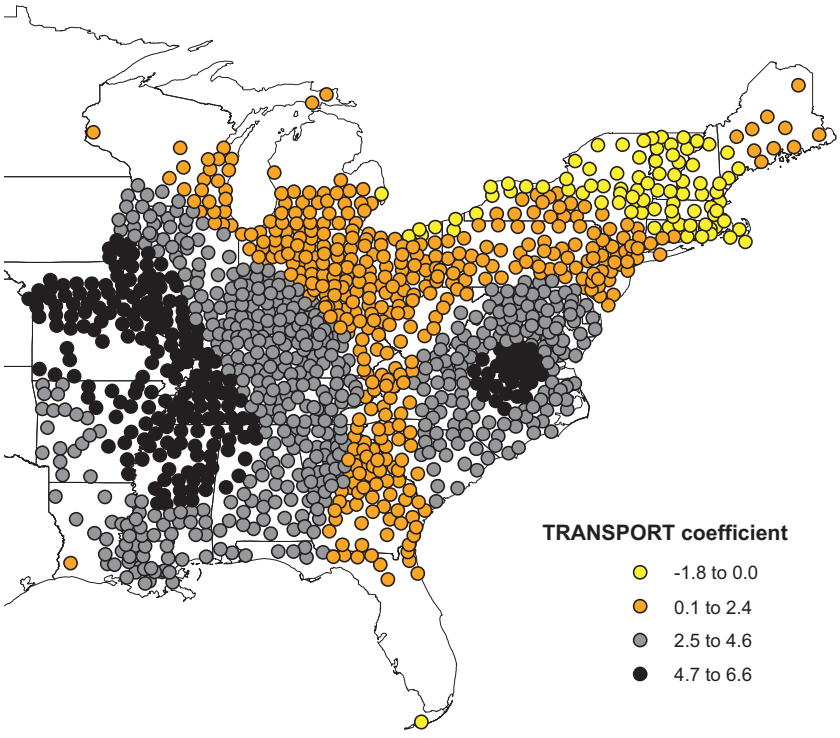

Fig. 3. The effect of transportation on crude death rates in the Antebellum United States.

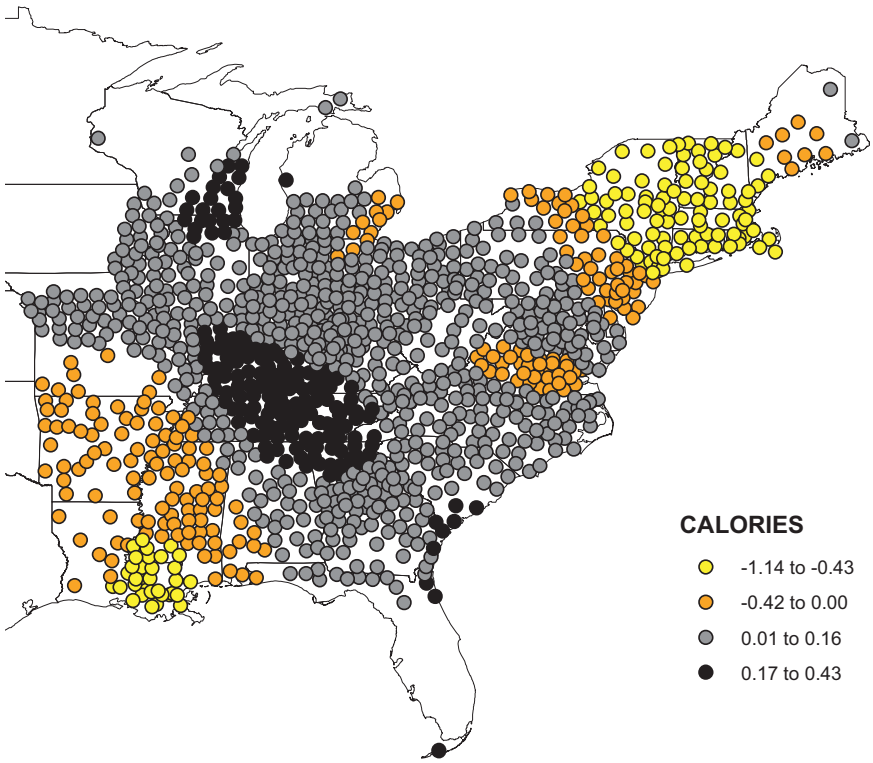

Fig. 4. The effect of calories on crude death rates in the Antebellum United States.

estimated risk of malaria in a county in the 1850s (Hong, 2007).

The *R*-squared of the GWR regression compared to OLS increases from 0.12 to 0.39 (GWR) or by a factor of 3.25.[9] (Because every location has its own coefficient for all the independent variables, we only report *R*-squared of the GWR.) Instead, 1251 coefficients of TRANSPORT independent variable are mapped in Fig. 3.

The GWR specification indicates that, with there was considerable spatial variation (Fig. 3). The mean value of the 1251 TRANSPORTATION coefficients is 2.87. Although transportation seems to have increased crude death rates in the GWR model on average as also found in the OLS model, the mean is misleading by itself, because it hides the spatial variation. We find that while the coefficients of TRANSPORTATION for the Midwestern farming area are highly positive, especially around the Mississippi, the Missouri, and the Ohio River basins, the coefficients for the industrialized Northeast tend to be negative. This implies that access to water transportation was associated with an increase in crude death rates in the Midwest, but with a decrease in crude death rates in the industrialized Northeast. One can infer from this the nutrient flows were associated with the CDR.

The spatial distribution of CALORIES coefficients on crude death rates (Fig. 4) parallels those of TRANSPORT, and provides an explanation for the anomaly of the OLS models estimated by Haines et al. (2003). Those models estimated that the effect of the calories surplus variable on crude death rates was positive. However, the regional pattern of CALORIES coefficients shows that a greater marketable calories surplus was associated with an increase in the crude death rate in the Midwest, but that it was associated with a decrease in the crude death rate in the Northeast. In other words, marketable calories surpluses worked to the health advantage of people in the Northeast but to a disadvantage to those in the Midwest. Farmers in the Midwest sold their calories surplus, but people in the Northeast consumed not only locally produced calories but also the amounts they purchased from the Midwest.

Similarly, I first run OLS regressions with individual height as the dependent variable thereby reproducing the Haines et al. (2003) results (Table 2 column 1).

Then I map the mean of residuals at each county (thereby reducing the 6557 estimates to the 550 counties) in Fig. 5. The county-level mean of the residuals are obviously not distributed randomly as the OLS model implicitly assumes, implying that geography was an important determinant of heights.

Then, I expand Haines et al.'s (2003) OLS model on heights[10] to a GWR specification in which all of the

[9] The increase of *R*-squared is a good signal, but it should be interpreted with a caveat. In GWR, u_m is defined by $u_m = y_m - x' b(m)$. Although, the definition of *R*-squared of GWR is comparable, it is not identical to the OLS *R*-squared (in OLS u_m is defined by $u_m = y_m - x' b$).

[10] The height regression model is slightly different from Haines et al. (2003). In this paper, I added MALARIA and dropped three insignificant year dummies ($YEAR1862_m$, $YEAR1863_m$, and $YEAR1864_m$) because time dummies may not have spatial variation.

Table 2
OLS regression results (heights of Union Army Recruit).

Dependent variable: height	(1)	(2)	(3)
Intercept	65.51*** (0.12)	68.40*** (0.14)	68.78*** (0.17)
Variable for individual recruits			
MOVER	0.11 (0.07)	0.10 (0.07)	0.10 (0.07)
YEAR1865	−0.47*** (0.14)	−0.47*** (0.14)	−0.48*** (0.14)
FARMER	0.48*** (0.07)	0.47*** (0.07)	0.50*** (0.10)
FARMER × NORTHEAST	–	–	−0.10 (0.13)
LABORER	0.11 (0.13)	0.12 (0.13)	0.17 (0.13)
Variable from the county in which the recruit was born			
URBAN	−1.23*** (0.16)	−1.21*** (0.16)	−1.12*** (0.16)
MALARIA	–	−0.51 (0.37)	−1.13** (0.50)
NORTHEAST	–	–	−0.54** (0.17)
TRANSPORT	−0.26*** (0.07)	−0.25*** (0.07)	−0.26*** (0.09)
TRANSPORT × NORTHEAST	–	–	0.19 (0.15)
CALORIES	0.02 (0.01)	0.02 (0.01)	0.01 (0.01)
WEALTH	−0.17*** (0.04)	−0.15*** (0.04)	−0.11** (0.04)
Adjusted *R*-squared	0.042	0.041	0.044
F	36.57***	32.32***	26.34***
N	6564	6557	6557

** Significant at 5% level.
*** Significant at 1% level.

coefficients in the equation are indexed by m ($m = 1,\ldots,6557$; 6557 is the number of observations).[11]

$$\begin{aligned}\text{HEIGHT}_m = {} & a(m) + b_1(m)\ \text{MOVER}_m \\ & + b_2(m)\,\text{YEAR}\,1865_m + b_3(m)\ \text{FARMER}_m \\ & + b_4(m)\ \text{LABORER}_m + b_5(m)\ \text{CALORIES}_m \\ & + b_6(m)\ \text{WEALTH}_m \\ & + b_7(m)\ \text{TRANSPORT}_m + b_8(m)\ \text{URBAN}_m \\ & + b_9(m)\ \text{MALARIA}_m + e_m \end{aligned} \quad (2)$$

where HEIGHT_m is the height in inches of the mth Union Army recruit. MOVER_m is a dummy variable, which takes one if the recruit enlisted in a county other than the one in which he was born, zero otherwise. YEAR1865_m is one if the mth recruit enlisted in 1865 when minimum height requirement was reduced, zero otherwise. FARMER_m and LABORER_m equals one if the individual was a farmer or laborer, and zero otherwise. The remaining variables are the same as those discussed in the mortality regression.

Applying the GWR model increases the *R*-squared of the regression increases from 0.0416 (OLS) to 0.0536 (GWR) or by 29%. The 6557 coefficients estimated and mapped for the transport variable (TRANSPORT) are again reduced to 550 values, by calculating average per county of birth (Fig. 6).

The GWR specification indicates that, with there was considerable spatial variation (Fig. 6). The mean value of the 6557 TRANSPORTATION coefficients is −0.17, but with considerable variation across regions. Hence, on average transportation decreased height in the GWR model, as in the OLS model. However, the mean is again quite misleading by itself because, while the coefficients for the Midwestern farming area are highly negative, especially around the Mississippi, the Missouri, and the Ohio River basins and the Erie Canal, the coefficients for the industrialized Northeast tend to be positive or slightly negative. This implies that access to water transportation was associated with a decrease in height in the Midwest, but with an increase in height in the industrialized Northeast as we expected on theoretical considerations.

The distribution of the coefficient on FARMER also has a very clear pattern (Fig. 7). On average farmers are half inch taller than non-farmers, but there is sharp regional pattern whereby farmers around urban centers near the coast are taller only by 0.25 in. whereas farmers in the more remote sections of upstate New York and Western Pennsylvania are taller by as much as 0.64 in. Indeed, the pattern is divided by the Appalachian Mountains. This result parallels the conclusion of Cuff (2005) who investigated Pennsylvania and argued that the Appalachian Mountains worked as a barrier for commercial farming. The GWR expands this result to the national level.

4. Comparison of OLS and GWR

The GWR results show that the effect of transportation, calories, and being a farmer differed greatly across regions. However, the regional effects of those variables cannot be captured in detail by an OLS model, not even by using regional dummies[12] and interaction terms.

Regional dummies (NORTHEAST, MIDWEST and SOUTH; column 3 of Table 1) and interaction terms (columns 4–5 of Table 1; column 3 of Table 2) improve OLS regression results, confirm GWR results, but provide only a rudimentary view of the true pattern as depicted in GWR (Figs. 3 and 5–7).

Insofar as OLS with regional dummy and interaction terms treats the whole region as homogeneous, it is unable

[11] Because the risk of malaria was unavailable in 9 counties, 6557 observations are used in the estimation.

[12] The state codes in 1850 clearly divide the three regions: Northeast (state code 1–19), Midwest (state code 20–39), and South (state code 40–59).

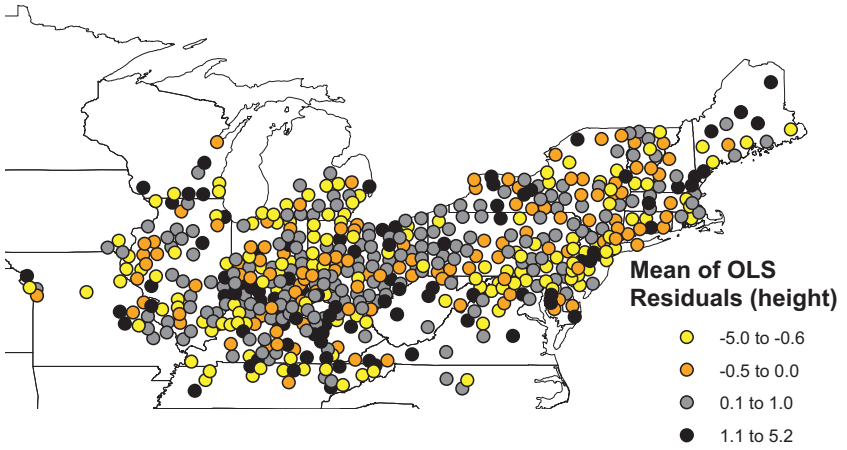

Fig. 5. Residuals of baseline OLS regression with height as dependent variable.

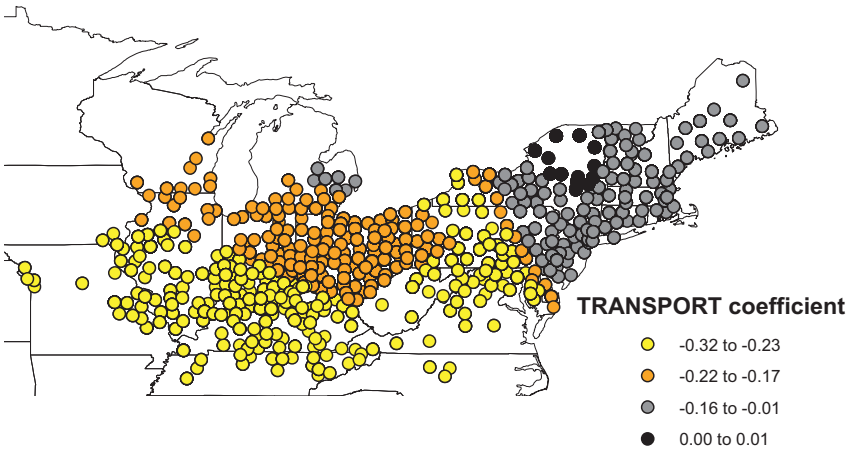

Fig. 6. The effect of transportation on height in the Antebellum United States.

to capture the spatial heterogeneity across the U.S. that depends so much on topography. Especially, the robust effects of transportation around the major rivers remain unnoticed using the OLS estimates. In other words, GWR yields a much finer grained resolution of spatial effects than do the dummy variables specification. This is the case, because the GWR specification is equivalent to running some 1251 regressions for the CDR model and 6557 regressions for the height model. Furthermore, it should be noted that without knowing GWR results specifying a model with the regional dummy variables[13] and interaction terms is not practical.

GWR also provides a plausible explanation for the positive coefficient on WEALTH. In the CDR OLS regression, more wealth increases mortality (Table 1). Haines et al. (2003) suggested inequality as a possible cause of the paradox. Sunder (2007) shows quite convincingly, however, that high income protected the rich from the adverse effects of modern economic growth. The spatial distribution of WEALTH coefficients (Fig. 8) provides a more plausible explanation for the health–wealth paradox. In fact, taking spatial effects into consideration indicates that WEALTH does not increase the crude death rate significantly in the North at all, but does so in the South. This pattern can be explained by an omitted variable bias insofar as slaves, one of the most important wealth assets of the South (Yang, 1984), were not included in the wealth measures of the OLS regressions of Haines et al. (2003).[14] Consequently, WEALTH of the Southern counties is underestimated. This systematic measurement problem

[13] Without interaction terms, the coefficient of MIDWEST and SOUTH are positive and significant (Table 1 column 3), meaning that living in the Midwest or South increased the crude death rate. This result is quite difficult to interpret without knowing the GWR results. Researchers are likely to expect that living in the Northeast increased the crude death rate because it was more urbanized, but the OLS regression shows the opposite. Unable to discern a refined spatial pattern using OLS, researchers might conclude that geography was irrelevant and exclude the regional dummies from the model altogether.

[14] WEALTH: Value of agricultural land, live-stock, and implements and manufacturing capital.

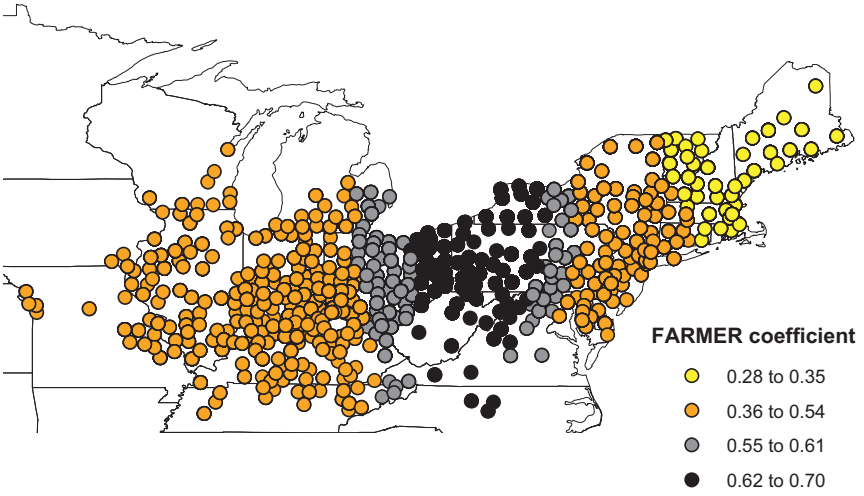

Fig. 7. The effect of being farmer on height in the Antebellum United States.

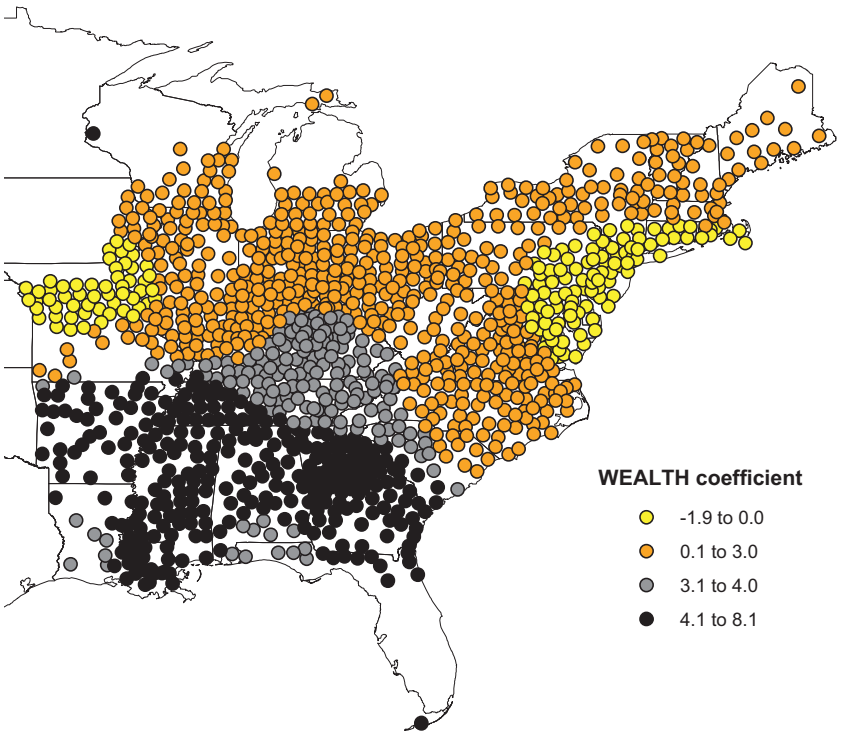

Fig. 8. The effect of wealth on crude death rates in the Antebellum United States.

can be detected by including a dummy variable for slavery (SOUTH) and an interaction term SOUTH × WEALTH. Table 1 column 5 shows that the health–wealth paradox exists mainly in the South.[15]

[15] The model estimates that WEALTH is associated with an increase in the crude death rates in the SOUTH by 2.02 (2.91/100 × 69.56: the average of WEALTH in the South is $69.5674). Considering that SOUTH (a proxy for slavery) decreases the crude death rate by 9.57, the net effect of wealth in the South is negative (the difference between SOUTH and MIDWEST is 3.94), which is consistent with expectations.

5. Discussion

The relation between transportation and commercial farming is well established and leads to a straightforward explanation for the spatial patterns in Figs. 3 and 6. The expansion of steamboats on the Mississippi River after 1815 reduced transportation costs to Midwestern farmers by a large amount. Before steamboats, the Mississippi River operated as a one-way water route. The cost of downstream transportation was low, but upstream transport rates were very high. Steamboats solved the problem of

upstream transportation. As a result, Midwestern farmers engaged in two-way trade with the Northeast (the Midwestern farmers traded their agricultural products for manufactured goods), and commercial farming was born. On net the two-way trade may have increased the income (and utility) of Midwestern farmers, but their families' nutritional status and life expectancy declined. In contrast, the Northeast could import food products at a lower cost than before, which led to improved nutrition. Hence, the geographic pattern of TRANSPORT coefficients supports the hypothesis first suggested in Komlos (1987).

The application of GWR methodology to the "Antebellum Puzzle" indicates that the Antebellum U.S. economy was not homogenous across regions and illustrates two major benefits of the method. First, as a form an exploratory data analysis, GWR is useful for measuring spatial relationships. GWR provides spatial disaggregation that helps to identify differences across space and local exceptions. A comparison of the OLS results using regional dummies and those of GWR shows the advantages of the latter insofar as the OLS dummy variable approach provides a superficial view of the spatial pattern. The GWR, in contrast, is able to capture spatial patterns using a much finer resolution such as the strong effects along the major rivers.

Second, GWR is useful for correcting specification errors in estimates where other explanatory variables are the center of interest. Although the regional differences in the antebellum U.S. have been recognized for a long time, researchers have not had such a clear representation of the regional differences in height and crude death rates as we are able to generate here with the use of maps. This is partly because modeling regional differences in OLS requires strong assumptions on the definition of regions. Even though the division of regions may be broadly correct, other parameters of interest suffer from omitted variable biases when spatial effects at the local level are excluded from the specification. GWR offers a spatial model with very few prior assumptions on the regional effects, which helps to correct specification errors.

6. Conclusion

Many social scientists are interested in measuring and interpreting spatial patterns. Traditional analysis, using OLS regression, implicitly assumes that the effect of the explanatory variables is constant across space. GWR enables researchers to relax this assumption and offers a deeper geographic perspective.[16]

The paradox of increasing income at a time when both physical stature and life expectancy were declining might be explained in cities by the urban penalty as, for instance, Haines (1998) showed for New York City in this period. However, the height decline also occurred in rural areas and was largest among farmers and did not affect the wealthy (Sunder, 2011; Lang and Sunder, 2003). The GWR results suggest that the force driving the puzzle in rural areas was primarily commercial farming as suggested first in Komlos (1987). The empirical evidence from GWR indicates that the access to transportation lowered nutritional status in the Midwest along the major rivers, but it enhanced nutritional status in the Northeast, as one would expect from the theory of interregional trade. The spatial pattern of coefficients for food surpluses is consistent with the notion that the transportation revolution changed the strategies of Midwestern farmers from self-sufficiency to production for distant markets, with adverse nutritional consequences for the children of the local population.

The analysis is based on the fact that Midwestern farmers were tall before the arrival of cheap water transportation. After the introduction of steamboats and the opening of the Erie Canal, Midwestern farmers with access to water transportation could exchange their products for manufactured goods. Not only water transportation, but also other methods such as wagons, canals, and railroads improved in the early nineteenth century, increasing the possibilities of commercial farming and integrating distant markets.

This transportation revolution delivered not only food to the Northeastern and foreign markets, but also created new economic opportunities for Midwestern farmers. The new possibilities provided by the transportation revolution integrated distant markets. This does not mean that the children of the commercial farmers did not have sufficient food to eat but that their mix of food changed markedly: they ate less meat and dairy products and more carbohydrates. It is more in line with the evidence that self-sufficient farmers had exceptional nutrition status at the start of the period under consideration. The marginal product of labor in farming in the U.S. was higher than in Europe, because of vast quantities of good land. Without easy access to markets, self-sufficient farmers faced low prices for their products and consumed most of what they produced.

"What did the commercially-minded farmers do with their cash earnings?" Komlos and Coclanis (1997) suggest that farmers bought more manufactured goods. However, the amount of nutrient-rich food consumed by the family declined. The following excerpt from a Willa Cather short story is illuminating.

> They had been at one accord not to hurry through life, not to be always skimping and saving. They saw their neighbours buy more land and feed more stock than they did, without discontent. Once when the creamery agent came to the Rosickys to persuade them to sell him their cream, he told them how much money the Fasslers, their nearest neighbours, had made on their cream last year.
>
> 'Yes,' said Mary, 'and look at them Fassler children! Pale, pinched little things, they look like skimmed milk. I'd rather put some colour into my children's faces than put money into the bank.' [Willa Cather, 'Neighbour Rosicky' from Cuff (2005)]

Thus, commercialization of farming had many positive aspects (particularly in the long run), such as stimulating

[16] In some sense, GWR estimates geographical heteroskedasticity using non-parametric methods.

investments in drainage or housing. On net the adults were undoubtedly better off with the trade, and the receiving side benefited as we showed above. Yet, there were some hidden negative aspects to commercialization for the children, particularly in the short and medium run, which should not be overlooked and is difficult to detect by other indicators besides anthropometry. The Antebellum Puzzle is to some extent a question of timing. The nutrients were gone immediately, while the farm investments might have had a positive impact but with a considerable time lag, and the intensity of the impact may not have sufficed to compensate for the loss of nutrients to the farmer's family.

On the other hand, the New England and mid-Atlantic states ended up with improved nutrition due to the transportation revolution. Although cheap transport delivered more nutritional surpluses, the Northeast underwent urbanization with its attendant deleterious effects on health and mortality. The geographic patterns of TRANSPORT and PROTEIN coefficients combined with the effect of urbanization support this conclusion. It should be noted that international trade (the export of corn and flour abroad) which was fostered by the transportation revolution also contributed to the Antebellum Puzzle. In fact, as much as 16–51% of corn and 20–41% of flour transported South on the Mississippi was exported abroad through New Orleans (Lindstrom, 1970).

Economists and economic historians may find troubling the idea that market integration in the nineteenth century could have hidden adverse consequences on the welfare of some parts of the population experiencing the onset of modern economic growth. For some time they have celebrated steamboats, canals and railroads as engines of prosperity and rising living standards and a core ingredient in industrialization. On the other hand, it is well known that "the standard of living" has many components, and while it is fully legitimate to celebrate the role of these innovations in improving material conditions, they also had negative externalities particularly for children that should not be neglected. From this perspective, the health revolution of the late nineteenth century was even more important and dramatic, eventually liberating the public from the health costs of trade. To be sure, as Sunder (2011) has shown, the wealthy were immune to these adverse developments insofar as they were able to pay the increased price of nutritional products.

This paper revisits the analysis of the Antebellum Puzzle by using anew technique, GWR, that estimates regional effects at the county level. Considering that many countries experienced and continue to experience regional differences in social and economic activities, the GWR methodology can be applied to many social and economic research problems. The fact that Chinese and Indian farmers trade their agricultural products for manufactured products suggests that those two large developing countries may experience the same or similar problem in their development paths. Considering that the commercialization can hurt the nutritional status of the succeeding generation, more attention should be paid to the study hidden burdens of economic development (Cuff, 2005).

References

A'Hearn, B., 1999. The Antebellum Puzzle revisited: a new look at the stature of Union Army Recruits during the Civil War. In: Komlos, J., Baten, J. (Eds.), The Biological Standard of Living in Comparative Perspective, Vol. I, The Americans, Asia, and Australia. Steiner Verlag, Stuttgart.

Bodenhorn, H., Rockoff, H., 1992. Regional interest rates in Antebellum America. In: Goldin, C., Rockoff, H. (Eds.), Strategic Factors in Nineteenth Century American Economic History. University of Chicago Press, Chicago.

Carter, S.B., Ransom, R.L., Sutch, R., 2004. Family matters: the life-cycle transition and the Antebellum American fertility decline. In: Guinnane, T.W., Sundstrom, W.A., Whatley, W.C. (Eds.), History Matters. Stanford University Press, Stanford.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74 (368), 829–836.

Craig, L., Weiss, T., 1998. Nutritional status and agricultural surpluses in the Antebellum United States. In: Komlos, J., Baten, J. (Eds.), Studies on the Biological Standard of Living in Comparative Perspectives. Franz Steiner Verlag, Stuttgart.

Cuff, T., 2005. The Hidden Cost of Economic Development. Ashgate, Burlington.

Easterlin, R.A., 1961. Regional income trends, 1840–1950. In: Harris, H. (Ed.), American Economic History. McGraw-Hill, New York.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester.

Goldin, C., Katz, L., 1999. The shaping of higher education: the formative years in the United States, 1890–1940. Journal of Economic Perspectives 13, 37–62.

Griliches, Z., 1960. Hybrid corm and the economics of innovation. Science 132 (3422), 275–280.

Haines, M., 1998. Health, height, nutrition, and mortality: evidence on the "Antebellum Puzzle" From Union Army Recruits for New York State and the United States. In: Komlos, J., Baten, J. (Eds.), The Biological Standard of Living in Comparative Perspective. Steiner, Stuttgard.

Haines, M., Craig, L., Weiss, T., 2003. The short and the dead: nutrition, mortality and the "Antebellum Puzzle" in the United States. Journal of Economic History 63 (2), 382–413.

Haines, M., Craig, L., Weiss, T., 2011. Did African Americans experience the 'Antebellum Puzzle'? Evidence from the United States Colored Troops during the Civil War. Economics and Human Biology 9 (1), 45–55.

Hong, S.C., 2007. Burden of early exposure to malaria in the United States, 1850–1860: malnutrition and immune disorders. Journal of Economic History 2007 (1), 1001–1035.

Komlos, J., 1998. Shrinking in a growing economy?. The mystery of physical stature during the industrial revolution. Journal of Economic History 58, 779–802.

Komlos, J., 1987. Height and weight of west point cadets: dietary change in Antebellum America. Journal of Economic History 47 (4), 897–927.

Komlos, J., Coclanis, P., 1997. On the puzzling cycle in the biological standard of living: the case of Antebellum Georgia. Explorations in Economic History 34 (4), 433–459.

Lang, S., Sunder, M., 2003. Non-parametric regression with Bayes X: a flexible estimation of trends in human physical stature in 19th century America. Economics and Human Biology 1 (1), 77–89.

Lindstrom, D., 1970. Southern dependence upon interregional grain supplies: a review of the trade flows 1840–1860. Agricultural History 44 (1), 101–113.

Steckel, R.H., 1995. Stature and the standard of living. Journal of Economic Literature 33, 1903–1940.

Sunder, M., 2007. Passports and economic development: an antropometric history of the U.S. Elite in the nineteenth century. Dissertation, University of Munich.

Sunder, M., 2011. Upward and onward: high-society American women eluded the antebellum puzzle. Economics and Human Biology 9, 165–171.

Wang, N., Mei, C.L., Yan, X.D., 2008. Local linear estimation of spatially varying coefficient models: an improvement on the geographically weighted regression technique. Environment and Planning A 40 (4), 986–1005.

Yang, D., 1984. Notes on the wealth distribution of farm households in the United States, 1860: a new look at two manuscript census samples. Explorations in Economic History 21, 88–102.

Zehetmayer, M., 2011. The continuation of the antebellum puzzle: stature in the US, 1847–1894. European Review of Economic History 15, 313–327.