GeoHealth

RESEARCH ARTICLE

10.1029/2020GH000320

Key Points:

- The significant spatial clusters of diabetes prevalence exist in the coastal northeast and northwest counties in Shandong Province, China
- A methodological improvement combined data-driven and spatial methods is proposed to identify automatically significant indicators
- Several economic, sociodemographic, education, and geographic environment indicators are found to be associated with diabetes prevalence

Correspondence to:

T. Fei and J. Wang,
feiteng@whu.edu.cn;
wangjian993@whu.edu.cn

Citation:

Li, Y., Fei, T., Wang, J., Nicholas, S., Li, J., Xu, L., et al. (2021). Influencing indicators and spatial variation of diabetes mellitus prevalence in Shandong, China: A framework for using data-driven and spatial methods. *GeoHealth*, *5*, e2020GH000320. https://doi.org/10.1029/2020GH000320

Received 1 SEP 2020
Accepted 22 FEB 2021

Influencing Indicators and Spatial Variation of Diabetes Mellitus Prevalence in Shandong, China: A Framework for Using Data-Driven and Spatial Methods

Yizhuo Li[1], Teng Fei[1], Jian Wang[2], Stephen Nicholas[3,4,5], Jun Li[1], Lizheng Xu[6], Yanran Huang[6], and Hanqi Li[1]

[1]School of Resource and Environmental Sciences, Wuhan University, Wuhan, China, [2]Research Center of Health Economics and Management, Dong Fureng Institute of Economic and Social Development, Wuhan University, Beijing, China, [3]Top Education Institute, Sydney, NSW, Australia, [4]Newcastle Business School, University of Newcastle, Newcastle, NSW, Australia, [5]School of Management and School of Economics, Tianjin Normal University, Tianjin, China, [6]School of Public Health, Center for Health Economics Experiment and Public Policy, Shandong University, Key Laboratory of Health Economics and Policy Research, NHFPC (Shandong University), Jinan, China

Abstract To control and prevent the risk of diabetes, diabetes studies have identified the need to better understand and evaluate the associations between influencing indicators and the prevalence of diabetes. One constraint has been that influencing indicators have been selected mainly based on subjective judgment and tested using traditional statistical modeling methods. We proposed a framework new to diabetes studies using data-driven and spatial methods to identify the most significant influential determinants of diabetes automatically and estimated their relationships. We used data from diabetes mellitus patients' health insurance records in Shandong province, China, and collected influencing indicators of diabetes prevalence at the county level in the sociodemographic, economic, education, and geographical environment domains. We specified a framework to identify automatically the most influential determinants of diabetes, and then established the relationship between these selected influencing indicators and diabetes prevalence. Our autocorrelation results showed that the diabetes prevalence in 12 Shandong cities was significantly clustered (Moran's $I = 0.328$, $p < 0.01$). In total, 17 significant influencing indicators were selected by executing binary linear regressions and lasso regressions. The spatial error regressions in different subgroups were subject to different diabetes indicators. Some positive indicators existed significantly like per capita fruit production and other indicators correlated with diabetes prevalence negatively like the proportion of green space. Diabetes prevalence was mainly subjected to the joint effects of influencing indicators. This framework can help public health officials to inform the implementation of improved treatment and policies to attenuate diabetes diseases.

1. Introduction

Diabetes mellitus is a common chronic disease caused by metabolic disorders. It has become one of the most frequent and widespread diseases in both developed and developing countries, associated with the epidemiology of obesity (Jia, Xue, Yin, et al., 2019) and its acute complications, such as diabetic ketoacidosis, diabetic coma, heart disease, and stroke (Dales et al., 2012). Besides its impact on diabetes suffers' health, diabetes imposes a large financial burden on individuals and a significant economic cost on a nation's health system, making diabetes one of the foremost public health challenges of the 21st century (Rong et al., 2016). According to the International Diabetes Federation (Domingueti et al., 2016), there are 425 million people with diabetes, accounting for 9% of the adult population worldwide, two thirds of whom are of working age. This figure is projected to reach 592 million globally by 2035, with 62.6 million in China, the country with the second-highest number of people diagnosed with diabetes (Xu et al., 2013).

Diabetes studies have shown that age, sex, educational level, regional economic development, and medical facilities level are all influencing indicators in the risk of diabetes (Brown et al., 2004; Hipp & Chalise, 2015; Maier et al., 2013). More recently, interdisciplinary collaboration, particularly with geography and sociology, and greater data accessibility offer diabetes studies a mass of meteorological and geo-environmental data on potential influencing indicators related to diabetes prevalence. Geographic information system (GIS)

and remote sensing methods (Jia et al., 2017) provide more abundant and effective geo-environmental data, such as the built environment (Feng et al., 2010), the accessibility of local food (Salois, 2012), urbanization degree (Cherubini et al., 1999; Zhou, Astell-Burt, Yin, et al., 2015), and physical activities (Hu et al., 1999). For example, the food environment, such as the density of restaurants, retail food stores, and supermarkets, can play an important role in the prevalence of diabetes through a population's diet (Jia, Xue, Cheng, & Wang, 2019; Salois, 2012). Pilot GIS studies have identified significant relationships between air pollution, including $PM_{2.5}$, SO_2, and NO_2 concentrations obtained from remote sensing images, and diabetes prevalence (Dales et al., 2012; Eze et al., 2015; Thiering & Heinrich, 2015). Land use types and land use mix have been related to diabetes and public health more generally (Christian et al., 2011; Su et al., 2016). Drawing on sociology, sociodemographic indicators, such as low socioeconomic status (Walker et al., 2011), family income (Dinca-Panaitescu et al., 2011), housing inequality (Wan & Su, 2016), education level (Zhou, Astell-Burt, Bi, et al., 2015), and immigrant status (Siordia et al., 2012), are potential diabetes influencing indicators. Diabetes has been associated with a population's socioeconomic inequities, frequently conceptualized as social deprivation (Connolly et al., 2000; Maier et al., 2013; Tompkins et al., 2010). Some diet and nutrition indicators, like fruits, vegetables, and cereals intake, have also been linked to diabetes (Ezzati & Riboli, 2013).

Understanding the association between these influencing indicators and the prevalence of diabetes can control and prevent the risk of diabetes-related illnesses, informing treatment regimens and guiding effective policies. But understanding these associations has faced two challenges. First, selected influencing indicators have been based mainly on subjective judgment or advice from expert consultants. Health problems are subject to complex interactions among different influencing indicators that are not equally important for diabetes prevalence in different areas. If the number of modeled influencing indicators does not correspond to the real factors, model accuracy will be negatively impacted (Frank & Friedman, 1993). One solution is to use data-driven methods to extract significant influencing indicators automatically. With unprecedented volumes of data now available, machine learning methods, such as ridge regression and lasso regression, have been applied widely in high dimensional feature selection (Reichstein et al., 2019). Surprisingly, the applications of these new data-driven methods as ways of identifying influencing indicators of diabetes prevalence have been largely ignored in diabetes research.

The second challenge is that public health results tend to be spatially clustered, and obvious regional differences exist in epidemiological disease prevalence (Chalkias et al., 2013). Traditional statistical modeling methods, such as ordinary least squares (OLS) regression (Green et al., 2003), logistic multilevel binomial regression (Jia, Xue, Cheng, & Wang, 2019; Maier et al., 2013), robust regression (Salois, 2012), and generalized linear models with natural cubic splines (Dales et al., 2012) have not addressed spatial issues. These conventional statistical methods are often based on the assumption that the samples are independent of each other, which contradicts data with spatial autocorrelation, where two measurements taken from geographically close locations are often more similar than measurements from a distant location. Thus, it is necessary to add spatial effects to traditional statistical models (LeSage & Pace, 2009). With the development of GIS and spatial analysis, new spatial techniques and methods have been proposed to explain the spatial association between diabetes and various influencing indicators (Shi & Wang, 2015). For example, spatial clustering patterns of diabetes prevalence have been detected and quantified (Tompkins et al., 2010), and spatial effects have been added to statistical models such as spatial regression models (Sridharan et al., 2007; Wan & Su, 2016; Weng et al., 2017) and geographical weighted regression models (Hipp & Chalise, 2015; Siordia et al., 2012).

New to diabetes research, this paper proposes and develops a step-by-step framework that combines a machine learning model and a spatial regression model. Using medical insurance data from Shandong province, China, we apply this framework to identify automatically the probable indicators that could influence diabetes prevalence significantly, and then, establish the relationship between these selected influencing indicators and diabetes prevalence considering the spatial autocorrelation effects. Our framework provides public health officials and urban planners with a new perspective to inform the implementation of improved treatment and policies to attenuate diabetes diseases.

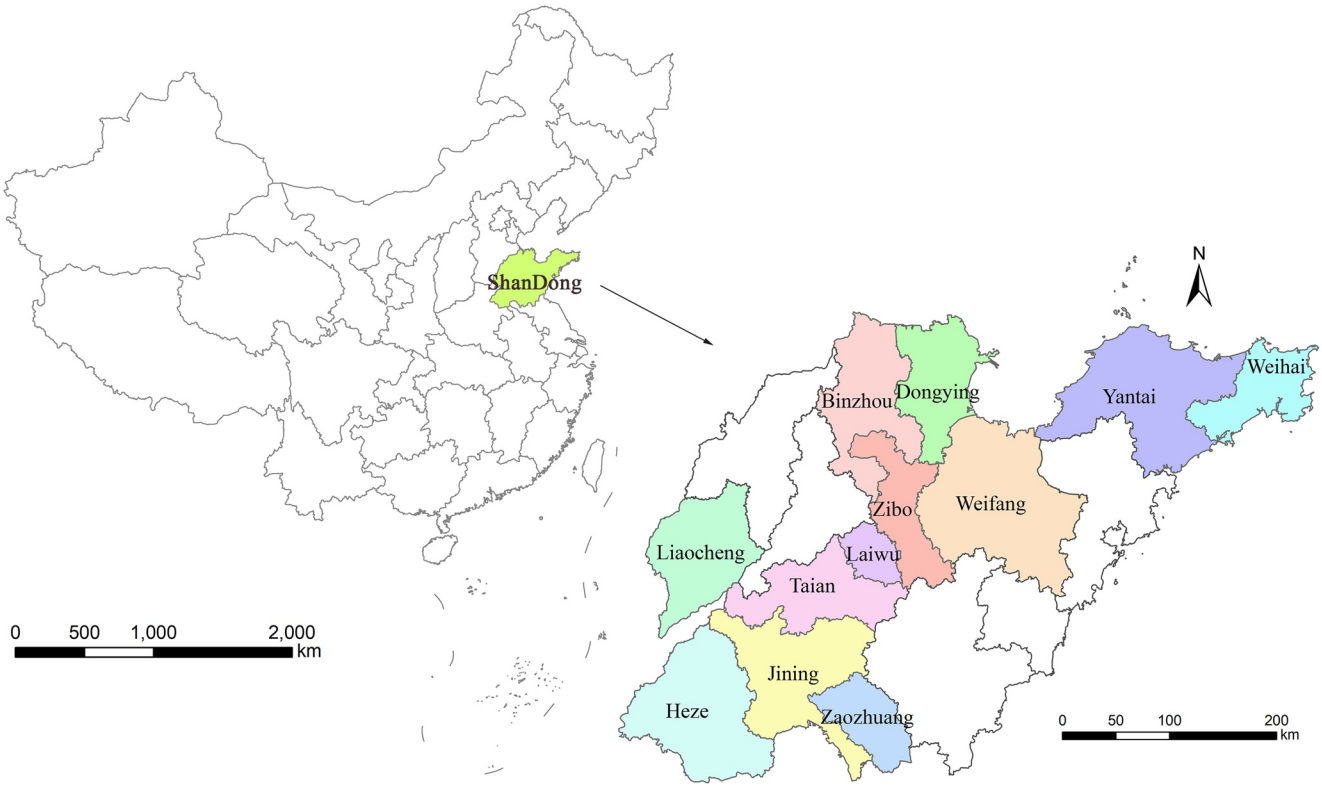

Figure 1. Location and administrative division of the study region in China.

2. Materials and Methods

2.1. Study Region

As shown in Figure 1, Shandong is an advanced industrial province on China's east coast and the lower reaches of the Yellow River between 34°25′ and 38°23′ north latitude and between 114°36′ and 122°43′ east longitude. With a population exceeding 100 million, Shandong had a gross domestic product (GDP) of RMB7.27 trillion (US$1.08 trillion) in 2017, ranking the third highest in China. As a largely industrial and fast-growing province, the total industrial added value was RMB2.87 trillion (US$425.07 billion) in 2017, an increase of 6.6% over the previous year, and an added value of agriculture of RMB280.2 billion (US$41.5 billion), an increase of 4.6% over the previous year. Shandong province has a warm temperate monsoon climate with four seasons. Its average annual temperature is 13°C Celsius and its rainfall is focused in the summer with 550–950 mm annual rainfall.

2.2. Diabetes Data

The type 2 diabetes mellitus data were obtained from a management database of patients' medical insurance in 12 Shandong cities in 2017, comprising 89 counties. The data involved inpatient and outpatient records with patient ID, sex, age, residential address at the county level, diagnosis results from the International Classification of Diseases (ICD) codes, hospital name, medical expenses, and medical insurance type. The medical insurance types consist of Urban Employed Basic Medical Insurance (UEBMI) and Integration of Urban and Rural Medical Insurance (IURMI), forming the basic medical insurance systems. Since there have been significantly different individual characteristics between these two types (Huang et al., 2019), it is worth exploring the patterns of diabetes by subgroups. China's basic medical insurance plans cover 95% of the population, thus the information from the medical insurance database can account for almost all the patients during the study period in each city. Based on the diagnosis results (ICD-10: E11–E14) recorded in the insurance database, diabetes mellitus patients were identified, and their inpatient and outpatient visit

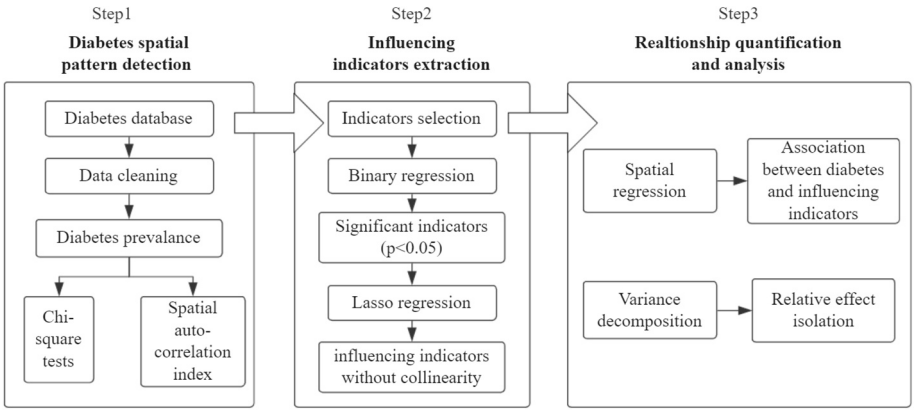

Figure 2. The methodological flowchart of our framework.

information were extracted including non-insulin-dependent diabetes mellitus, diabetic complications, and diabetic comorbidities. Due to the lack of patients' accurate addresses, the annual diabetes prevalence was calculated at the county level.

2.3. Influencing Indicators

We collected 46 influencing indicators at the county level, categorized into four different domains: economic, sociodemographic, education, and geographical environment. For parsimony and space limitations, we only listed the important influencing factors after indicators extraction mentioned in Section 2.4.2: (1) Economic indicators included per capita total export and per capita GDP, which were collected from the 2017 Shandong statistical yearbook; (2) sociodemographic indicators included per capita retail sales of social consumer goods, per capita grain production, per capita fruit production, and per capita meat production, which were also obtained from the 2017 statistical yearbook; (3) educational indicators included per capita teacher amounts in general secondary school and per capita teacher amounts in primary school from the statistical yearbook; (4) geographical environmental indicators were derived and calculated based on GIS tools in the county scale. A spatial database was created in which GIS layers of spatial environmental data were recorded in ArcGIS version 10.2 (ESRI Inc.) software. Specifically, the average elevation for each county was calculated from digital elevation model. Based on the Kriging interpolation and zonal statistic methods, the annual average sunshine hours were obtained from meteorological stations in the China Meteorological Data Network, and the annual average $PM_{2.5}$ were calculated from air quality monitoring stations in Shandong Province. The proportions of land use types were calculated by implementing supervision classification in the ENVI software (Exelis Visual Information Solutions Company) on the remote sensing images, downloaded from Google Earth. Road network density in each county was counted and calculated based on the road length of the main roads (for any form of motor transport) and secondary roads (supplementing a main road at moderate or slow speeds), which were crawled from Amap (https://lbs.amap.com/api/javascript-api/reference/layer#TileLayer.RoadNet). The accessibilities of medical facilities were calculated from the number of medical facilities within 1 h of walking buffers from residences based on the medical facilities and residence Points of Interest (POI) in 2017.

2.4. Framework

The flowchart of our framework is shown in Figure 2. The procedure includes three steps. First, the Chi-square tests were used to analyze the statistical significance of stratification differences and the autocorrelation index was applied to detect spatial patterns of diabetes. If the spatial clustering pattern exists in diabetes of prevalence, it is necessary to use the spatial regression model. Second, before analyzing the relationship between the indicators and diabetes prevalence, we used binary linear regression and lasso regression to extract the most significant influencing indicators of diabetes without collinearity. Lastly, we applied the spatial regression model to analyze how the spatial prevalence of diabetes is associated with

significant diabetes correlates and utilized variance decomposition to isolate the relative effect of influencing indicators on diabetes prevalence.

2.4.1. Step 1: Data Processing and Spatial Clustering Analysis

Diabetes data were divided into subgroups according to inpatient or outpatient record types and medical insurance types. After filtering and removing duplicate visiting records based on the patient ID, the number of diabetes patients in each county was calculated. A Chi-square test was used to analyze the significance of the difference in these subgroups at the county level, which was done by SPSS 22.0 software (SPSS Inc.). If the difference was significant, these subgroups were chosen to proceed to the next step. Five subgroups were formed: Group 1 Inpatients; Group 2 Outpatients; Group 3 UEBMI Patients; Group 4 IURMI Patients; and Group 5 All Patients in our database. To normalize disease values, the diabetes prevalence rates of five subgroups in each county were expressed as (Coggon et al., 1997)

$$p_i = \frac{n_i}{N} \tag{1}$$

where p_i is the prevalence rate for diabetes in county i; n_i is the number of patients who suffered from diabetes in county i with a specific period of time; N is the total population in each county collected from the 2017 statistical yearbook. The spatial distributions of diabetes prevalence were visualized by a series of thematic maps, which were classified into six classes using the "natural breaks" method by seeking to minimize each class's average deviation from the class mean, while maximizing each class's deviation from the means of the other classes (Faka et al., 2017).

To detect global spatial clustering patterns and quantify the degree of spatial autocorrelation of diabetes prevalence, Moran's I index was applied (Moran, 1950). Moran's I index ranges between -1 and $+1$, where the value close to $+1$ indicates a strong spatial correlation of the diabetes prevalence rate, while -1 shows spatial dispersal. To find the location of significantly similar or dissimilar clustering, local spatial autocorrelation is quantified by calculating the Local Moran's I index (Anselin, 1995). Positive values indicate spatial clustering of similar values, and negative values show the clustering of dissimilar values. The outcomes of Local Moran's I include five possible categories to identify the existence of pockets or clusters: High-High, Low-Low, High-Low, Low-High, and Not Significant. Analyses were done with GeoDa software, version 1.12 using queen contiguity weights, and the parameter of the order contiguity was set to 1.

2.4.2. Step 2: Influencing Indicators Extraction

In Step 2, two analyses were conducted to identify the most essential influencing indicators without multicollinearity. First, binary linear regression was performed between each indicator and diabetes prevalence. Indicators that were statistically significant ($p < 0.05$) were selected. Second, to eliminate the multicollinearity problem, lasso regression was utilized on these selected indicators. The lasso regression extracts prognostic signatures from large databases driven entirely by the data itself. It uses the absolute coefficient function of the model as a penalty term to compress the coefficients of the model, achieving the purpose of variable selection and parameter estimation simultaneously (Hayes et al., 2015; Tibshirani, 1996). By weighing the deviation variance of the model, the lasso regression overcomes the shortcoming of traditional indicator selection methods, like stepwise multiple regression, and effectively maintains the interpretability of selected variables that have explicit properties (Guo et al., 2015; Mueller-Using et al., 2016). In our study, cross-validation with 10 folds was used to tune the regularization parameter. The influencing indicators with non-zero coefficients in the sparsest model were then chosen into Step 3. Analyses were done with R software, version 3.3.2.

2.4.3. Step 3a: Spatial Regression

The spatial autocorrelation phenomenon has always existed in public health studies (Chalkias et al., 2013; Wan & Su, 2016), and it violates the independence assumption of errors in OLS. Thus, in Step 3, we applied spatial regression modeling to analyze the associations between diabetes prevalence and the significant influencing indicators in different subgroups incorporating the spatial autocorrelation dependency. There are two commonly used spatial regression models: the Spatial Lag regression in Equation 2 and the Spatial Error regression in Equation 3 (Anselin, 2013).

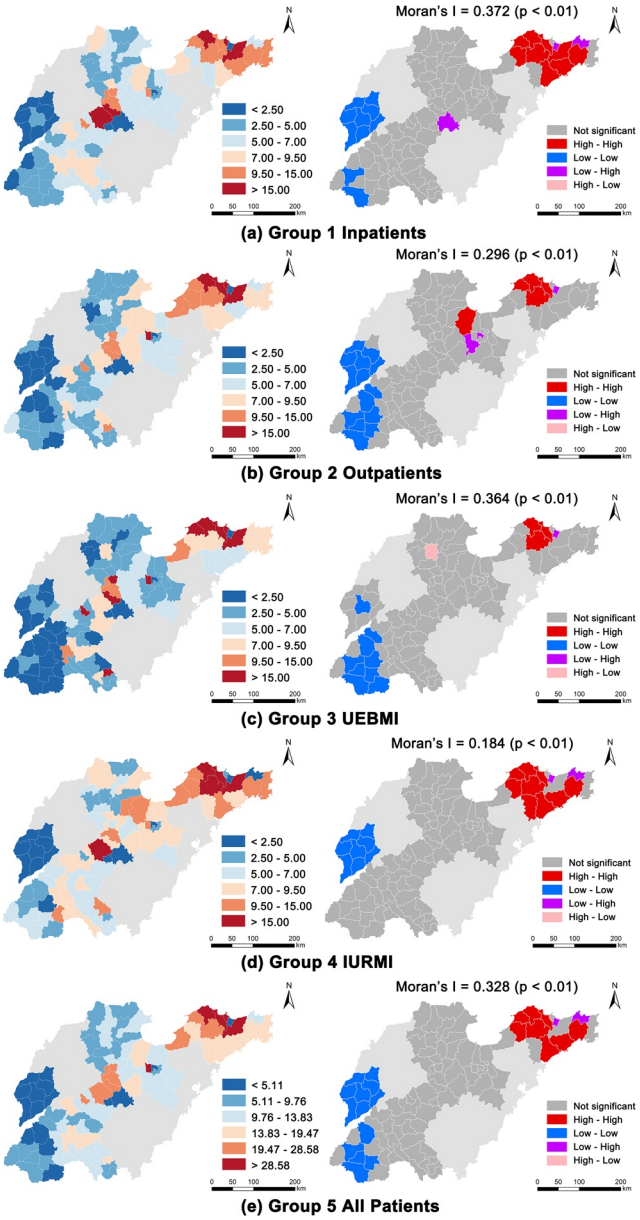

Figure 3. The spatial patterns of diabetes prevalence rates per 1,000 people in 12 Shandong cities at the county level based on different subgroups with the global Moran' *I* index and LISA maps.

$$Y = \alpha + \beta X + \lambda W_Y + e \tag{2}$$

$$Y = \alpha + \beta X + e\left(e = \lambda W_e + u\right) \tag{3}$$

where X is influencing indicators; Y is the diabetes prevalence rate for each county; β is the coefficients for each indicator; e is the error term; W_Y is the spatial weight matrix for the dependent variable and W_e is the spatial weight matrix for error term; λ is the spatial autoregressive coefficient; α and u are the scalar variables.

Robust Lagrange Multiplier (LM) tests were used to determine the type of spatial regression by judging which LM value in the regression models was more significant illustrated in LeSage and Pace (2009). All the indicators were normalized before modeling. Spatial regression modeling was implemented in the GeoDa software.

2.4.4. Step 3b: Variances Decomposition

To compare the relative importance of essential influencing indicators on the diabetes prevalence rate, we employed the variance decomposition (VD) method (Anderson & Cribble, 1998; Su et al., 2014). VD can decompose the variances of the dependent variable into shares and compare the relative effect of different exploratory variables by calculating individual or joined effects (Heikkinen et al., 2005). We classified the significant influencing indicators into four categories as illustrated before. Next, the total explained variance was calculated and decomposed into several sections: (1) the individual effects of four categories of influencing indicators; (2) the joint effect of two categories of influencing indicators; (3) the joint effect of three categories of influencing indicators; and (4) the joint effect of four categories of influencing indicators.

3. Results

The number of diabetes patients was 387,954, 371,617, 356,044, and 403,527 in four subgroups (inpatient, outpatient, UEBMI, and IURMI) and 759,571 in All Patients Group, respectively. The Chi-square test results indicated that different subgroups displayed significant differences at the county level, comprising inpatient or outpatient (Chi-square = 34882.868, $p < 0.001$) and patients using UEBMI or IURMI (Chi-square = 114966.648, $p < 0.001$). Thus, it is necessary to implement the grouping process in subsequent analysis.

For the five groups, Figure 3 showed the diabetes prevalence rates in each county in the left subfigures and their global Moran' *I* index values. Diabetes prevalence per 1,000 people in 12 Shandong cities at the county level ranged from 0.97% to 49.43% for all patients (Group 5) and was significantly clustered (Moran's $I = 0.328$, $p < 0.01$). The spatial distribution patterns in Figure 3 were broadly similar for the different groups. Generally, the coastal north-eastern counties presented a higher prevalence of diabetes, while those in the western part, especially the northwest region, exhibited lower diabetes prevalence. Some differences in spatial distributions were also found, with UEBMI patients (Group 3) displaying lower diabetes prevalence than IURMI patients (Group 4).

Heterogeneity of diabetes prevalence was observed among the counties from the global Moran's *I* and local autocorrelation results. Counties in the north-eastern region have High-High values presenting spatial ag-

Table 1
Associations Between Diabetes Prevalence and Influencing Indicators at the County Scale in 2017 in 12 Shandong Cities, China

Domain	Significant influencing indicators	Group 1 - Inpatients	Group 2 - Outpatients	Group 3 - UEBMI	Group 4 - IURMI	Group 5 - All
ECO	Per capita total export	–	13.736***	7.948*	–	–
	Per capita GDP	0.992	–	3.921***	–	–
SOC	Per capita retail sales of social consumer goods	5.912**	6.830***	–	15.916***	14.356***
	Per capita grain production	−1.347	−2.796	–	−6.492***	−9.486**
	Per capita fruit production	8.500***	7.052**	10.972***	–	17.773***
	Per capita meat production	–	–	2.032	–	–
EDU	Per capita teacher amounts in general secondary school	3.609*	–	–	–	7.167
GEO	Average elevation	1.619	–	–	–	−3.874
	Proportion of building land	0.931	–	2.618	–	–
	Proportion of green space	–	−5.786***	−0.763	−2.095	−3.914
	Proportion of blue space	−1.849	−5.595***	−4.486**	–	−8.882***
	Proportion of bare soil land	–	0.805	–	3.182	−0.583
	Annual average sunshine hours	–	–	1.173	–	–
	Annual average $PM_{2.5}$	−0.120	1.938	1.290	–	1.382
	Road density	–	–	−6.000**	–	–
	The accessibility of hospital	–	–	−1.568	–	–
	The accessibility of clinic	–	–	–	–	−2.739
Constant		3.207*	7.559***	4.340	4.972**	14.062***
R^2		0.539	0.612***	0.535	0.610***	0.629***

Abbreviations: ECO, economic level; SOC, sociodemographic level; EDU, education level; GEO, geographical environment level.
*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

gregation phenomena (red color), while counties in the western region have Low-Low prevalence clustering patterns (blue color). These results confirm that a spatial regression method should be applied to analyze the relationship between diabetes prevalence and influencing indicators.

Before applying the spatial regression model, we extracted 29 influencing indicators that were statistically significant ($p < 0.05$) after binary linear regression, and then selected 17 influencing indicators after lasso regression, which became final determinants of diabetes (Table 1). Note that different subgroups may include different significant indicators.

Based on these selected indicators, the results of the robust LM test in spatial regression indicated that all LM Error values in five groups were more significant ($p < 0.01$) than LM Lag ($p > 0.01$). Thus, the Spatial Error model with the queen weight matrix, incorporating the average neighboring influences in the geographical space, was chosen as the final model.

Table 1 displayed the association between different significant influencing indicators and diabetes prevalence based on the Spatial Error model. Two economic indicators including per capita total export and per capita GDP promoted the increase of diabetes prevalence. Sales of social consumer goods played an essential role in shaping public health, with per capita retail sales of social consumer goods a significant positive influence on diabetes prevalence across all groups except UEBMI. Unexpectantly, per capita grain production presented a negative correlation with diabetes prevalence (Group 4 IURMI and Group 5 All Patients) although some groups were insignificant (Group 1 Inpatients and Group 2 Outpatients). Per capita fruit production showed a significant positive relationship with diabetes prevalence across all groups. The education indicator had a significant positive relationship with diabetes prevalence in Group 1 Inpatients, but it was insignificant in Group 5 All Patients, which was contrary to our expectations. All the land use type indicators were extracted. The proportion of green space (e.g., vegetation and woodland) and blue space

Table 2
The Individual and Joint Effects of Different Categories of Influencing Indicators in Terms of Their Contributions to the Total Variations on Diabetes Prevalence (%)

Domain	Group 1 - Inpatients	Group 2 - Outpatients	Group 3 - UEBMI	Group 4 - IURMI	Group 5 - All
ECO	9.92	**22.33**	2.81	–	13.59
SOC	10.69	5.12	3.09	4.29	6.41
EDU	1.15	–	–	–	–
GEO	–	5.12	**44.94**	**61.90**	1.55
ECO & SOC	1.15	0.93	–	3.81	3.88
ECO & EDU	17.37	–	–	–	13.79
ECO & GEO	–	8.37	**23.60**	0.48	–
SOC & GEO	**18.89**	**20.47**	**15.73**	**13.33**	**21.17**
EDU & GEO	2.48	–	–	–	1.17
ECO & SOC & GEO	–	**37.67**	9.83	**16.19**	**20.19**
ECO & EDU & GEO	1.72	–	–	–	2.14
SOC & EDU & GEO	**18.70**	–	–	–	–
ECO & SOC & EDU & GEO	**17.94**	–	–	–	**16.12**
Total	100	100	100	100	100

Note. Bold numbers denote the top three largest proportions.
Abbreviations: ECO, economic level; SOC, sociodemographic level; EDU, education level; GEO, geographical environment level.

(e.g., water and wetlands) presented significant negative relationships with diabetes prevalence across all groups except in the Inpatients and IURMI. High $PM_{2.5}$ concentration increased the risk of diabetes. High accessibility of medical facilities presented a negative correlation in Group 3 UEBMI and Group 5 All Patients, though they were not significant.

Contributions of these significant influencing indicators to the total variations of diabetes prevalence in each group are displayed in Table 2. For Group 3 and Group 4, the individual effects of the geographic environment were stronger than those of the other categories, which indicated that UEBMI and IURMI patients were more influenced by these environmental indicators. The educational factors contributed less to the total variations individually. Besides, the joint effects between sociodemographic and geographical environment indicators, as well as those between economic with sociodemographic and geographical environment indicators accounted for a relatively high proportion of the total explained variances. In Group 1 Inpatients and Group 5 All Patients, the joint influences between all four categories of factors were also relatively strong. Such results suggested that in most cases, the joint effect of influencing indicators explained more diabetes prevalence.

4. Discussion

This paper proposed a framework to extract the influencing indicators automatically and to evaluate the associations between the most essential influencing indicators and diabetes prevalence, with data-driven and spatial methods new to diabetes research. We found that these influencing indicators and estimated coefficients varied across different groups, which revealed complex associations and potential mechanisms between diverse indicators and diabetes prevalence though they did not specify causal relationships.

In Figure 3, we observed spatial clustering patterns of diabetes prevalence among the counties in Shandong Province. The degree of spatial autocorrelation of diabetes prevalence was significant (Moran's $I = 0.328$, $p < 0.01$). This result was consistent with previous diabetes studies (Green et al., 2003; Hipp & Chalise, 2015; Siordia et al., 2012; Tompkins et al., 2010). Although diabetes is a health threat all over the world, its prevalence and distribution in various areas are heterogeneous. Also, we found that the coastal north-eastern counties in both Inpatients and Outpatients groups existed several clusters of high prevalence of diabetes.

These areas have been far from the city center and socially and ethnically diverse with high socioeconomic deprivation, which may cause this hotspot phenomenon. The specific reasons need further analysis. There were also different diabetes prevalence and spatial distribution patterns in different subgroups such as UEBMI and IURMI. UEBMI patients had lower diabetes prevalence than IURMI patients. It may be subject to average individual characteristics in different groups (Huang et al., 2019). IURMI patients mainly consisted of rural residents, urban retired, unemployed, students, and children with lower education levels and poor physical conditions, which may increase the prevalence of diabetes.

We found that several economic indicators had a positive correlation with diabetes prevalence. With rapid urbanization and the province's economic transition, risks of diseases have increased, negatively influencing public health and lifestyle (Gong et al., 2012). Previous studies have shown that diabetes prevalence has been greatly affected by economic factors, and socioeconomic inequalities will result in higher risks to diabetes (Fano et al., 2012; Grintsova et al., 2014). From our results, socioeconomic factors, including per capita retail sales of social consumer goods, per capita total export and per capita GDP, were significantly positively associated with the diabetes prevalence (Couchoud et al., 2011; Li et al., 2018). A previous study on a regional scale also suggested that rapid economic growth and transition may pose a potential higher risk of diabetes prevalence compared to low GDP regions (Tang et al., 2019). Possible explanations for these findings are that the urban areas with high GDP attract poor rural migratory workers, who have poorer health, lack health education and are less health-conscious. Moreover, people in high economic development areas tend to consume high-calorie foods, which may lead to obesity, one of the main causes of diabetes (Kastorini & Panagiotakos, 2009; Tang et al., 2019).

We also found an association between food production and the prevalence of diabetes. Interestingly, our results indicated that fruit production had a significant and positive relationship associated with diabetes prevalence, while grain production moderated the prevalence. This observation seems to agree with a global study based on country-level data demonstrating that fruit intake significantly positively influenced the prevalence of diabetes, and cereal intake had a negative association (Li et al., 2020). Although the production of food cannot totally stand for residents' volume of food intake, we speculate that there are two potential pathways concerning the effect of food production on diabetes prevalence. First, grain production may represent the availability of adequate food supply for farmers. Some researchers have identified inadequate food supply as a potential risk factor for diabetes, so-called food deserts (Maier et al., 2013; Seligman et al., 2009), which may bring higher diabetes prevalence. Second, excessive fructose intake (>50 g/d) may be one of the underlying pathogens of type 2 diabetes (Johnson et al., 2007, 2009), though fructose may not cause health problem (DiNicolantonio et al., 2015). In the case where fruit products cannot be completely sold or exported, we surmise that these fruits may be processed and consumed by local people and the fructose generated in this process may have a potential impact on health. It is interesting to relate industrialization level with this indicator in the future.

Higher education levels generally lead to a better understanding of healthcare instructions, such as glycemic control (Schillinger et al., 2002). However, our results indicated that the educational indicator was positively correlated with the diabetes prevalence though only one group was significant. One explanation might be that this indicator did not fully evaluate the education level of each county, especially urban counties with high levels of migrant workers. Some studies have also argued that the association between education level and diabetes prevalence has been uncertain and insignificant (Couchoud et al., 2011; Zhou, Astell-Burt, Bi, et al., 2015), which is consistent with our empirical results.

Many selected geographic environment indicators were observed affecting the associations with diabetes prevalence. Our results supported the general hypothesis that pleasant and comfortable environments positively influenced people's physical and mental health (Corman et al., 2016; Wan & Su, 2016). We found that the proportion of green space and blue space both had significantly negative relationships with diabetes prevalence, consistent with previous research (Faka et al., 2017; Li et al., 2020; Mackenbach et al., 2014). This is also supported by recent urban planning research linking land use types to people's lifestyles and health (Gascon et al., 2016; Su et al., 2016). They can be explained as follows: first, green space can purify the air environment (Alcock et al., 2015) and blue space is regarded as an important factor in absorbing pollutants and harmful substances (Völker & Kistemann, 2015), which may slow down the risk of diabetes prevalence; second, areas with a high proportion of green and recreational spaces motivate people to enhance the physical activity, social interaction, and emotional connections (Faka et al., 2017; Wall et al., 2012),

which moderate diabetes prevalence and provide plenty of public health benefits. In addition, exposure to $PM_{2.5}$ pollution has been regarded as a significant factor influencing the prevalence of diabetes (Coogan et al., 2016; Eze et al., 2017; Puett et al., 2011). We also found that weakly positive associations between $PM_{2.5}$ and diabetes prevalence, which was similar to several epidemiologic studies demonstrating air pollution increased the prevalence of diabetes and acute diabetic complications explained by subclinical inflammation (Dales et al., 2012; Krämer et al., 2010).

We observed that individual effects of economic factors on diabetes prevalence were significant, while other factors were more subjected to the joint effects, which was consistent with previous studies on the influence of social deprivation on public health (Wan & Su, 2016). It suggested that diabetes resulted from the interactions and combinations of multiple influencing indicators from different domains.

Some studies have suggested that cold temperature could lead to elevating glycosylated hemoglobin levels and acute complications of diabetes (Hou et al., 2017; Huang et al., 2019). Also, the proportion of the secondary and third industry output has often been used in related diabetes research (Couchoud et al., 2011). However, after executing the binary linear regression and lasso regression, these indicators were insignificant and were not included in the final spatial regression model.

Our methodological framework combined a data-driven method with a spatial regression model to improve the performance in estimating the association of indicators with diabetes prevalence. Two suggestions can be proposed for future directions of disease spatial modeling. First, it is necessary to perform factor pre-extraction as was done in our study; and, second, if spatial autocorrelation phenomenon exists, spatial components need to be considered by applying spatial analysis. The former guarantees the parsimony and adaptability of the modeling, while the latter improves the accuracy of the modeling. This framework can be easily extended to other regions or other diseases similar to diabetes associated with many uncertain indicators and drivers. Such selected influencing indicators promise powerful insights for the local health policy makers and medical practitioners to propose tailored advice and decision-making support to solve the issues in public health from the perspective of urban planning and economics. The influencing factors also help government policy makers to plan and regulate the external environment.

This study had some limitations. First, our analysis was based on counties as basic spatial units, which were identified according to administrative divisions. Importantly, in the health geography field, different study scales, such as the Thiessen polygon (Openshaw, 1984) or creating zones with certain characteristics (Li et al., 2019), may generate different results and statistical bias, the so-called modifiable area unit problem. Second, we did not consider the temporal series of indicators and the patients' exposures to risk factors for a long time. In this paper, we mainly focused on the differences among different counties. Lastly, our framework was exploratory and did not specify the causal mechanisms. To improve our framework, the next step would involve identifying the causal path and mediator variables by combing structural equation modeling (SEM) or mediating effect tests (Wan & Su, 2016).

5. Conclusion

In this study, we proposed a framework to measure the association between diabetes prevalence and influencing indicators with spatial effects. The significant influencing indicators were identified automatically using a data-driven method new to diabetes research. We not only detected the spatial patterns of diabetes prevalence in five different groups (inpatient, outpatient, UEBMI, IURMI, and All), but also isolated the individual and joint effects of influencing indicators on diabetes prevalence. We also performed a comprehensive exploration of the influence of economic, sociodemographic, education, and geographic environment indicators on diabetes prevalence. Finally, we provided detailed methodological improvements to help public health departments treat diabetes diseases and build healthy environments. This framework can be extended to other regions or other diseases to explore corresponding relationships between diseases and influencing indicators.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The data including diabetes prevalence data across each county and the elevation data, and the codes crawling the POIs of medical facilities and obtaining remote sensing images that support the findings of this study are available at doi.org/10.6084/m9.figshare.14061104. Specific diabetes mellitus patients' health insurance records are available through Huang (2019). The meteorology data are available in the China Meteorological Data Network and other data can be obtained from the 2017 Shandong statistical yearbook (https://www.chinayearbooks.com/shandong-statistical-yearbook-2017.html).

Acknowledgments
The authors would like to thank Shandong Medical Insurance Research Association to provide fund support under Award Number SK170078.

References

Alcock, I., White, M. P., Lovell, R., Higgins, S. L., Osborne, N. J., Husk, K., & Wheeler, B. W. (2015). What accounts for 'England's green and pleasant land? A panel data analysis of mental health and land cover types in rural England. *Landscape and Urban Planning*, *142*, 38–46. https://doi.org/10.1016/j.landurbplan.2015.05.008

Anderson, M. J., & Cribble, N. A. (1998). Partitioning the variation among spatial, temporal and environmental components in a multivariate data set. *Austral Ecology*, *23*(2), 158–167. https://doi.org/10.1111/j.1442-9993.1998.tb00713.x

Anselin, L. (1995). Local indicators of spatial association – LISA. *Geographical Analysis*, *27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Anselin, L. (2013). *Spatial econometrics: Methods and models (Vol. 4)*. Springer Science & Business Media. Retrieved from https://www.springer.com/gp/book/9789024737352

Brown, A. F., Ettner, S. L., Piette, J., Weinberger, M., Gregg, E., Shapiro, M. F., et al. (2004). Socioeconomic position and health among persons with diabetes mellitus: A conceptual framework and review of the literature. *Epidemiologic Reviews*, *26*(1), 63–77. https://doi.org/10.1093/epirev/mxh002

Chalkias, C., Papadopoulos, A. G., Kalogeropoulos, K., Tambalis, K., Psarra, G., & Sidossis, L. (2013). Geographical heterogeneity of the relationship between childhood obesity and socio-environmental status: Empirical evidence from Athens, Greece. *Applied Geography*, *37*, 34–43. https://doi.org/10.1016/j.apgeog.2012.10.007

Cherubini, V., Carle, F., Gesuita, R., Iannilli, A., Tuomilehto, J., Prisco, F., et al. (1999). Large incidence variation of Type I diabetes in central-southern Italy 1990-1995: Lower risk in rural areas. *Diabetologia*, *42*(7), 789–792. https://doi.org/10.1007/s001250051228

Christian, H. E., Bull, F. C., Middleton, N. J., Knuiman, M. W., Divitini, M. L., Hooper, P., et al. (2011). How important is the land use mix measure in understanding walking behavior? Results from the RESIDE study. *International Journal of Behavioral Nutrition and Physical Activity*, *8*(1), 55. https://doi.org/10.1186/1479-5868-8-55

Coggon, D., Rose, G., & Barker, D. (1997). Quantifying diseases in populations. In J. Critchley (Ed.), *Epidemiology for the uninitiated* (p. 4). Wiley.

Connolly, V., Unwin, N., Sherriff, P., Bilous, R., & Kelly, W. (2000). Diabetes prevalence and socioeconomic status: A population based study showing increased prevalence of type 2 diabetes mellitus in deprived areas. *Journal of Epidemiology & Community Health*, *54*(3), 173–177. https://doi.org/10.1136/jech.54.3.173

Coogan, P. F., White, L. F., Yu, J., Burnett, R. T., Seto, E., Brook, R. D., et al. (2016). PM2.5 and diabetes and hypertension incidence in the Black women's health study. *Epidemiology*, *27*(2), 202. https://doi.org/10.1097/EDE.0000000000000418

Corman, H., Curtis, M. A., Noonan, K., & Reichman, N. E. (2016). Maternal depression as a risk factor for children's inadequate housing conditions. *Social Science & Medicine*, *149*, 76–83. https://doi.org/10.1016/j.socscimed.2015.11.054

Couchoud, C., Guihenneuc, C., Bayer, F., Lemaitre, V., Brunet, P., & Stengel, B. (2011). Medical practice patterns and socio-economic factors may explain geographical variation of end-stage renal disease incidence. *Nephrology Dialysis Transplantation*, *27*(6), 2312–2322. https://doi.org/10.1093/ndt/gfr639

Dales, R. E., Cakmak, S., Vidal, C. B., & Rubio, M. A. (2012). Air pollution and hospitalization for acute complications of diabetes in Chile. *Environment International*, *46*, 1–5. https://doi.org/10.1016/j.envint.2012.05.002

Dinca-Panaitescu, S., Dinca-Panaitescu, M., Bryant, T., Daiski, I., Pilkington, B., & Raphael, D. (2011). Diabetes prevalence and income: Results of the Canadian Community Health Survey. *Health Policy*, *99*(2), 116–123. https://doi.org/10.1016/j.healthpol.2010.07.018

DiNicolantonio, J. J., O'Keefe, J. H., Lucan, S. C. (2015). Added fructose: A principal driver of type 2 diabetes mellitus and its consequences. *Mayo Clinic Proceedings*, *90*(3), 372–381. https://doi.org/10.1016/j.mayocp.2014.12.019

Domingueti, C. P., Dusse, L. M. S. A., Carvalho, M. d. G., de Sousa, L. P., Gomes, K. B., & Fernandes, A. P. (2016). Diabetes mellitus: The linkage between oxidative stress, inflammation, hypercoagulability and vascular complications. *Journal of Diabetes and Its Complications*, *30*(4), 738–745. https://doi.org/10.1016/j.jdiacomp.2015.12.018

Eze, I. C., Foraster, M., Schaffner, E., Vienneau, D., Héritier, H., Rudzik, F., et al. (2017). Long-term exposure to transportation noise and air pollution in relation to incident diabetes in the SAPALDIA study. *International Journal of Epidemiology*, *46*(4), 1115–1125. https://doi.org/10.1093/ije/dyx020

Eze, I. C., Hemkens, L. G., Bucher, H. C., Hoffmann, B., Schindler, C., Künzli, N., et al. (2015). Association between ambient air pollution and diabetes mellitus in Europe and North America: Systematic review and meta-analysis. *Environmental Health Perspectives*, *123*(5), 381–389. https://doi.org/10.1289/ehp.1307823

Ezzati, M., & Riboli, E. (2013). Behavioral and dietary risk factors for noncommunicable diseases. *New England Journal of Medicine*, *369*(10), 954–964. https://doi.org/10.1056/nejmra1203528

Faka, A., Chalkias, C., Montano, D., Georgousopoulou, E. N., Tripitsidis, A., Koloverou, E., et al. (2017). Association of socio-environmental determinants with diabetes prevalence in the Athens Metropolitan Area, Greece: A spatial analysis. *The Review of Diabetic Studies*, *14*(4), 381. https://doi.org/10.1900/rds.2017.14.381

Fano, V., Pezzotti, P., Gnavi, R., Bontempi, K., Miceli, M., Pagnozzi, E., et al. (2012). The role of socio-economic factors on prevalence and health outcomes of persons with diabetes in Rome, Italy. *The European Journal of Public Health*, *23*(6), 991–997. https://doi.org/10.1093/eurpub/cks168

Feng, J., Glass, T. A., Curriero, F. C., Stewart, W. F., & Schwartz, B. S. (2010). The built environment and obesity: A systematic review of the epidemiologic evidence. *Health & Place*, *16*(2), 175–190. https://doi.org/10.1016/j.healthplace.2009.09.008

Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), 109–135. https://doi.org/10.1080/00401706.1993.10485033

Gascon, M., Triguero-Mas, M., Martínez, D., Dadvand, P., Rojas-Rueda, D., Plasència, A., & Nieuwenhuijsen, M. J. (2016). Residential green spaces and mortality: A systematic review. *Environment International*, *86*, 60–67. https://doi.org/10.1016/j.envint.2015.10.013

Gong, P., Liang, S., Carlton, E. J., Jiang, Q., Wu, J., Wang, L., & Remais, J. V. (2012). Urbanization and health in China. *The Lancet*, *379*(9818), 843–852. https://doi.org/10.1016/s0140-6736(11)61878-3

Green, C., Hoppa, R. D., Young, T. K., & Blanchard, J. F. (2003). Geographic analysis of diabetes prevalence in an urban area. *Social Science & Medicine*, *57*(3), 551–560. https://doi.org/10.1016/s0277-9536(02)00380-5

Grintsova, O., Maier, W., & Mielck, A. (2014). Inequalities in health care among patients with type 2 diabetes by individual socio-economic status (SES) and regional deprivation: A systematic literature review. *International Journal for Equity in Health*, *13*(1), 43. https://doi.org/10.1186/1475-9276-13-43

Guo, P., Zeng, F., Hu, X., Zhang, D., Zhu, S., Deng, Y., & Hao, Y. (2015). Improved variable selection algorithm using a LASSO-type penalty, with an application to assessing hepatitis B infection relevant factors in community residents. *PLoS ONE*, *10*(7), e0134151. https://doi.org/10.1371/journal.pone.0134151

Hayes, J., Thygesen, H., Tumilson, C., Droop, A., Boissinot, M., Hughes, T. A., et al. (2015). Prediction of clinical outcome in glioblastoma using a biologically relevant nine-microRNA signature. *Molecular Oncology*, *9*(3), 704–714. https://doi.org/10.1016/j.molonc.2014.11.004

Heikkinen, R. K., Luoto, M., Kuussaari, M., & Pöyry, J. (2005). New insights into butterfly-environment relationships using partitioning methods. *Proceedings of the Royal Society B*, *272*(1577), 2203–2210. https://doi.org/10.1098/rspb.2005.3212

Hipp, J. A., & Chalise, N. (2015). Peer reviewed: Spatial analysis and correlates of county-level diabetes prevalence, 2009–2010. *Preventing Chronic Disease*, *12*, 140404. https://doi.org/10.5888/pcd12.140404

Hou, L., Li, M., Huang, X., Wang, L., Sun, P., Shi, R., et al. (2017). Seasonal variation of hemoglobin A1c levels in patients with type 2 diabetes. *International Journal of Diabetes in Developing Countries*, *37*(4), 432–436. https://doi.org/10.1007/s13410-016-0500-y

Hu, F. B., Sigal, R. J., Rich-Edwards, J. W., Colditz, G. A., Solomon, C. G., Willett, W. C., et al. (1999). Walking compared with vigorous physical activity and risk of type 2 diabetes in women. *Journal of American Medical Association*, *282*(15), 1433–1439. https://doi.org/10.1001/jama.282.15.1433

Huang, Y., Li, J., Hao, H., Xu, L., Nicholas, S., & Wang, J. (2019). Seasonal and monthly patterns, weekly variations, and the holiday effect of outpatient visits for type 2 diabetes mellitus patients in China. *International Journal of Environmental Research and Public Health*, *16*(15), 2653. https://doi.org/10.3390/ijerph16152653

Jia, P., Cheng, X., Xue, H., & Wang, Y. (2017). Applications of geographic information systems (GIS) data and methods in obesity-related research. *Obesity Reviews*, *18*(4), 400–411. https://doi.org/10.1111/obr.12495

Jia, P., Xue, H., Cheng, X., & Wang, Y. (2019). Effects of school neighborhood food environments on childhood obesity at multiple scales: A longitudinal kindergarten cohort study in the USA. *BioMed Central Medicine*, *17*(1), 99. https://doi.org/10.1186/s12916-019-1329-2

Jia, P., Xue, H., Yin, L., Stein, A., Wang, M., & Wang, Y. (2019). Spatial technologies in obesity research: Current applications and future promise. *Trends in Endocrinology and Metabolism*, *30*(3), 211–223. https://doi.org/10.1016/j.tem.2018.12.003

Johnson, M. T. J., Gersch, T. M., Segal, K. S., Feig, D. I., Lozada, J. L., Nakagawa, T., et al. (2007). Potential role of sugar (fructose) in the epidemic of hypertension, obesity and the metabolic syndrome. *Diabetes, Kidney Disease, and Cardiovascular Disease*, *86*(4), 899. https://doi.org/10.1093/ajcn/86.4.899

Johnson, R. J., Perez-Pozo, S. E., Sautin, Y. Y., Manitius, J., Sanchez-Lozada, L. G., Feig, D. I., et al. (2009). Hypothesis: Could excessive fructose intake and uric acid cause type 2 diabetes? *Endocrine Reviews*, *30*(1), 96–116. https://doi.org/10.1210/er.2008-0033

Kastorini, C.-M., & Panagiotakos, D. (2009). Dietary patterns and prevention of type 2 diabetes: From research to clinical practice; a systematic review. *Cancer Drug Resistance*, *5*(4), 221–227. https://doi.org/10.2174/157339909789804341

Krämer, U., Herder, C., Sugiri, D., Strassburger, K., Schikowski, T., Ranft, U., & Rathmann, W. (2010). Traffic-related air pollution and incident type 2 diabetes: Results from the SALIA cohort study. *Environmental Health Perspectives*, *118*(9), 1273–1279. https://doi.org/10.1289/ehp.0901689

LeSage, J., & Pace, R. K. (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC. Retrieved from https://www.routledge.com/Introduction-to-Spatial-Econometrics/LeSage-Pace/p/book/9781420064247

Li, J., Wang, S., Han, X., Zhang, G., Zhao, M., & Ma, L. (2020). Spatiotemporal trends and influence factors of global diabetes prevalence in recent years. *Social Science & Medicine*, *256*, 113062. https://doi.org/10.1016/j.socscimed.2020.113062

Li, L., Qiu, W., Xu, C., & Wang, J. (2018). A spatiotemporal mixed model to assess the influence of environmental and socioeconomic factors on the incidence of hand, foot and mouth disease. *BioMed Central Public Health*, *18*(1), 274. https://doi.org/10.1186/s12889-018-5169-3

Li, Y., Fei, T., & Zhang, F. (2019). A regionalization method for clustering and partitioning based on trajectories from NLP perspective. *International Journal of Geographical Information Science*, *33*(12), 2385–2405. https://doi.org/10.1080/13658816.2019.1643025

Mackenbach, J. D., Rutter, H., Compernolle, S., Glonti, K., Oppert, J.-M., Charreire, H., et al. (2014). Obesogenic environments: A systematic review of the association between the physical environment and adult weight status, the SPOTLIGHT project. *BMC Public Health*, *14*(1), 233. https://doi.org/10.1186/1471-2458-14-233

Maier, W., Holle, R., Hunger, M., Peters, A., Meisinger, C., Greiser, K. H., et al. (2013). The impact of regional deprivation and individual socio-economic status on the prevalence of Type 2 diabetes in Germany. A pooled analysis of five population-based studies. *Diabetic Medicine*, *30*(3), 78–86. https://doi.org/10.1111/dme.12062

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1/2), 17–23. https://doi.org/10.2307/2332142

Mueller-Using, S., Feldt, T., Sarfo, F. S., & Eberhardt, K. A. (2016). Factors associated with performing tuberculosis screening of HIV-positive patients in Ghana: LASSO-based predictor selection in a large public health data set. *BioMed Central Public Health*, *16*(1), 563. https://doi.org/10.1186/s12889-016-3239-y

Openshaw, S. (1984). *The modifiable areal unit problem*. UK: Geo books. Retrieved from https://ci.nii.ac.jp/naid/10003011548/

Puett, R. C., Hart, J. E., Schwartz, J., Hu, F. B., Liese, A. D., & Laden, F. (2011). Are particulate matter exposures associated with risk of type 2 diabetes? *Environmental Health Perspectives*, *119*(3), 384–389. https://doi.org/10.1289/ehp.1002344

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N., Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Rong, S., Le, C., Wenlong, C., Jianhui, H., Dingyun, Y., & Allison, G. (2016). Multilevel analysis of socioeconomic determinants on diabetes prevalence, awareness, treatment and self-management in ethnic minorities of Yunnan Province, China. *International Journal of Environmental Research and Public Health*, *13*(8), 751. https://doi.org/10.3390/ijerph13080751

Salois, M. J. (2012). Obesity and diabetes, the built environment, and the 'local' food economy in the United States, 2007. *Economics and Human Biology*, *10*(1), 35–42. https://doi.org/10.1016/j.ehb.2011.04.001

Schillinger, D., Grumbach, K., Piette, J., Wang, F., Osmond, D., Daher, C., et al. (2002). Association of health literacy with diabetes outcomes. *Journal of the American Medical Association*, *288*(4), 475–482. https://doi.org/10.1001/jama.288.4.475

Seligman, H. K., Laraia, B. A., & Kushel, M. B. (2009). Food insecurity is associated with chronic disease among low-income NHANES participants. *Journal of Nutrition*, *140*(2), 304–310. https://doi.org/10.3945/jn.109.112573

Shi, X., & Wang, S. (2015). Computational and data sciences for health-GIS. *Annals of Geographical Information Science*, *21*(2), 111–118. https://doi.org/10.1080/19475683.2015.1027735

Siordia, C., Saenz, J., & Saenz, J., Tom, S. E. (2012). An introduction to macro-level spatial nonstationarity: A geographically weighted regression analysis of diabetes and poverty. *Human Geography*, *6*(2), 5. https://doi.org/10.5719/hgeo.2012.62.5

Sridharan, S., Tunstall, H., Lawder, R., & Mitchell, R. (2007). An exploratory spatial data analysis approach to understanding the relationship between deprivation and mortality in Scotland. *Social Science & Medicine*, *65*(9), 1942–1952. https://doi.org/10.1016/j.socscimed.2007.05.052

Su, S., Hu, Y. N., Luo, F., & Mai, G., Wang, Y. (2014). Farmland fragmentation due to anthropogenic activity in rapidly developing region. *Agricultural Systems*, *131*, 87–93. https://doi.org/10.1016/j.agsy.2014.08.005

Su, S., Zhang, Q., Pi, J., Wan, C., & Weng, M. (2016). Public health in linkage to land use: Theoretical framework, empirical evidence, and critical implications for reconnecting health promotion to land use policy. *Land Use Policy*, *57*, 605–618. https://doi.org/10.1016/j.landusepol.2016.06.030

Tang, K., Wang, H., Liu, Y., & Tan, S. H. (2019). Interplay of regional economic development, income, gender and type 2 diabetes: Evidence from half a million Chinese. *Journal of Epidemiology & Community Health*, *73*(9), 867–873. https://doi.org/10.1136/jech-2018-211091

Thiering, E., & Heinrich, J. (2015). Epidemiology of air pollution and diabetes. *Trends in Endocrinology and Metabolism*, *26*(7), 384–394. https://doi.org/10.1016/j.tem.2015.05.002

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tompkins, J. W., Luginaah, I. N., Booth, G. L., & Harris, S. B. (2010). The geography of diabetes in London, Canada: The need for local level policy for prevention and management. *International Journal of Environmental Research and Public Health*, *7*(5), 2407–2422. https://doi.org/10.3390/ijerph7052407

Völker, S., & Kistemann, T. (2015). Developing the urban blue: Comparative health responses to blue and green urban open spaces in Germany. *Health & Place*, *35*, 196–205. https://doi.org/10.1016/j.healthplace.2014.10.015

Walker, J. J., Livingstone, S. J., Colhoun, H. M., Lindsay, R. S., McKnight, J. A., Morris, A. D., et al. (2011). Effect of socioeconomic status on mortality among people with type 2 diabetes: A study from the Scottish Diabetes Research Network Epidemiology Group. *Diabetes Care*, *34*(5), 1127–1132. https://doi.org/10.2337/dc10-1862

Wall, M. M., Larson, N. I., Forsyth, A., Van Riper, D. C., Graham, D. J., Story, M. T., & Neumark-Sztainer, D. (2012). Patterns of obesogenic neighborhood features and adolescent weight. *American Journal of Preventive Medicine*, *42*(5), 65–75. https://doi.org/10.1016/j.amepre.2012.02.009

Wan, C., & Su, S. (2016). Neighborhood housing deprivation and public health: Theoretical linkage, empirical evidence, and implications for urban planning. *Habitat International*, *57*, 11–23. https://doi.org/10.1016/j.habitatint.2016.06.010

Weng, M., Pi, J., Tan, B., Su, S., & Cai, Z. (2017). Area deprivation and liver cancer prevalence in Shenzhen, China: A spatial approach based on social indicators. *Social Indicators Research*, *133*(1), 317–332. https://doi.org/10.1007/s11205-016-1358-6

Xu, S., Ming, J., Xing, Y., Gao, B., Yang, C., Ji, Q., & Chen, G. (2013). Regional differences in diabetes prevalence and awareness between coastal and interior provinces in China: A population-based cross-sectional study. *BMC Public Health*, *13*(1), 299. https://doi.org/10.1186/1471-2458-13-299

Zhou, M., Astell-Burt, T., Bi, Y., Feng, X., Jiang, Y., Li, Y., et al. (2015). Geographical variation in diabetes prevalence and detection in China: Multilevel spatial analysis of 98,058 adults. *Diabetes Care*, *38*(1), 72–81. https://doi.org/10.2337/dc14-1100

Zhou, M., Astell-Burt, T., Yin, P., Feng, X., Page, A., Liu, Y., et al. (2015). Spatiotemporal variation in diabetes mortality in China: Multilevel evidence from 2006 and 2012. *BMC Public Health*, *15*(1), 633. https://doi.org/10.1186/s12889-015-1982-0