Bayesian Spatial and Temporal Machine Learning for Texas Diabetes Surveillance

Mohit Chhaparia, Lauren Pierce, Janghwan Suk, Leo Brown

Problem Statement

Goal of study

We will model county-level diabetes prevalence in Texas and assess spatial-temporal patterns within the response variable and the chosen covariates. In doing so, we will assess whether or not diabetes prevalence is spatially dependent (and potentially impacted by where in Texas a person lives) and if there was a significant change in diabetes prevalence over time for the years 2017 to 2021. By analyzing the chosen covariates, we will determine if the data reflects that certain factors are correlated with diabetes prevalence, which could allow for more accurate determination of risk or distribution of assistance to aid in prevention of this life-changing condition.

Data source

All non-spatial data for this project was obtained from the CDC's Diabetes Atlas[1] and spatial data for the counties was pulled from the U.S. Census Bureau's ACS regional database.

The non-spatial data consists of county-level estimates of diagnosed diabetes and obesity, gathered from CDC's Behavioral Risk Factor Surveillance System (BRFSS)[2] and from the U.S. Census Bureau's Population Estimates Program (PEP)[3]. The BRFSS is a telephone survey conducted of the U.S. population 18 and older, completed on a monthly basis within the state, but only those aged 20 and older were documented for estimation, to be consistent with the census information[2]. This data, unlike the data collected as a whole based on census information, is only that which can be collected by the survey respondents, and may have inherent selection bias. The data collection report states that those who responded to the question "Has a doctor ever told you that you have diabetes?" were included in the diagnosed diabetes group, with the exception of women who only reported this occurrence during pregnancy[2]. Obesity was defined as those with a BMI of 30 or higher, based on their self-reported height and weight. This does not account for muscle distribution, as has frequently been considered a contentious health measure.

Furthermore, the county-level estimates of the BRFSS, particularly when the county sample sizes were too small or not available, were generated with indirect model-dependent estimates using Small Area Estimate (SAE) techniques[4-6]. Within this data, we narrowed our scope to the counties of Texas, and, in addition to obesity percentage, considered the following county-level covariates of diagnosed diabetes for the years 2017 to 2021: whether a county was rural or urban[7], the percentage of children in poverty[8], and the percentage of people with no health insurance[9].

Methods applied

We initially applied Ordinary Least Squares Regression to narrow down the available covariates and get a preliminary overview of the interaction between diabetes prevalence and other factors. After applying the Moran Test to determine spatial dependency, we looked at Kriging and Spatial Temporal Gaussian Process Models, Geographically Weighted Regression (GWR), and nearest neighbor models including Spatial Nearest Neighbor Gaussian Process Model (spNNGP) and Spatial Conjugate Nearest Neighbor Gaussian Process Model (spConjNNGP), as well as related NNGP temporal models.

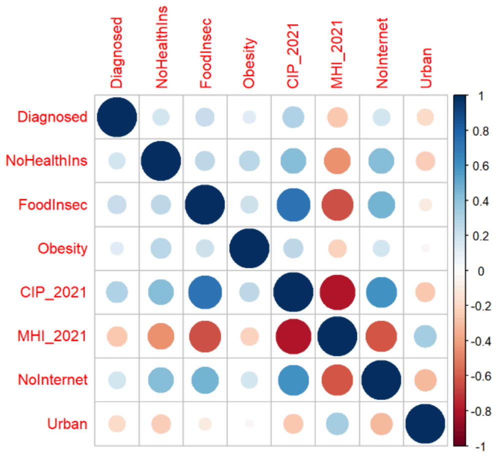
Data Summary

The working dataset contains 254 Texas counties with 9 variables. Four variables are stored as character IDs/labels (GEOID, county name, county label, and the spatial geometry), one variable (Urban) is a factor with two levels indicating whether a county is classified as urban or rural, and the remaining four variables are numeric health and socio-economic indicators.

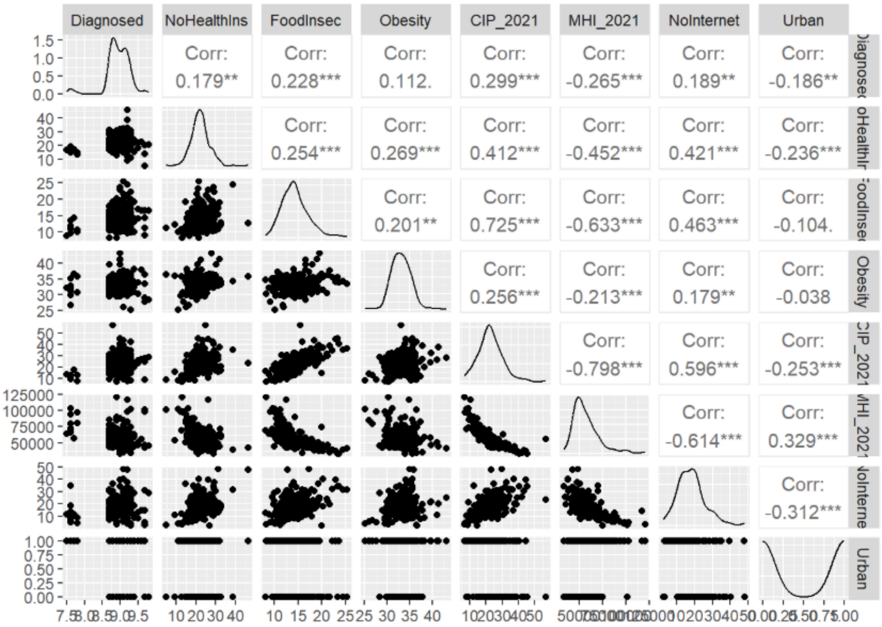
The summary shows that the dataset is complete with no missing values for any variable, so all 254 counties can be used in subsequent analyses without additional imputation or filtering. Among the numeric variables, diabetes diagnosis in 2021 (Diagnosed_2021) is homogeneous across counties (mean around 9% with a narrow range ~7.5–9.8). By contrast, no health insurance, obesity, and children-in-poverty percentages show much larger variation, indicating substantial heterogeneity in underlying risk factors across counties.

Exploratory Data Analysis

After eliminating potential covariates in the data set that had high levels of missingness for all years and low r^2 values, we were left with Food Environment Index, Food Insecurity, Median Income, Rural/Urban (0/1), Obesity (%), No Internet Access (%), Children in Poverty (%), and No Health Insurance (%). The Food Environment Index was excluded, as we had missing values for some counties, and the index included within it other covariates, which we wished to analyze separately. Internet Access was kept only for GWR, as that was not a health factor and could have selection bias for those who could answer the phone in the initial survey.



We then determined that there was significant (-0.798) inverse correlation between Median Household Income (MHI) and Children in Poverty (CIP) and with Food Insecurity (-0.633). Likewise, there was significant positive correlation between Children in Poverty and Food Insecurity (0.725). As these were all ways to determine a household's financial stability, we reduced these determinants to Children in Poverty, as it had the highest correlation with being diagnosed with diabetes.

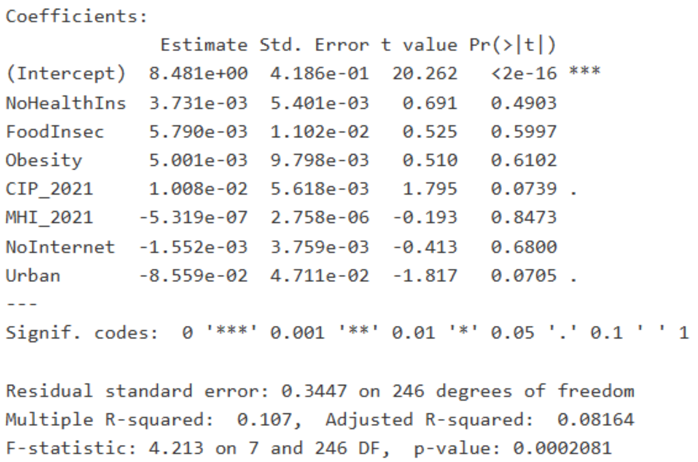


Data Modeling – Ordinary Least Squares Regression

As part of the Exploratory Data Analysis, Ordinary Least Squares Regression was applied to the response variable and the covariates. We compared the full model to various partial models and compared the AIC, BIC, and adjusted R^2. We also utilized Cook's Distance and standardized residuals to determine outliers and potential influence points within the models to see if any counties had an undue effect on the data.

Full Model

Diagnosed_Diabetes ~ No_Health_Insurance + Food_Insecurity + Obesity + CIP_2021 + MHI_2021 + No_Internet + Urban



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.481e+00  4.186e-01  20.262   <2e-16 ***
NoHealthIns  3.731e-03  5.401e-03   0.691   0.4903    
FoodInsec    5.790e-03  1.102e-02   0.525   0.5997    
Obesity      5.001e-03  9.798e-03   0.510   0.6102    
CIP_2021     1.008e-02  5.618e-03   1.795   0.0739 .  
MHI_2021    -5.319e-07  2.758e-06  -0.193   0.8473    
NoInternet  -1.552e-03  3.759e-03  -0.413   0.6800    
Urban       -8.559e-02  4.711e-02  -1.817   0.0705 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3447 on 246 degrees of freedom
Multiple R-squared:  0.107,	Adjusted R-squared:  0.08164 
F-statistic: 4.213 on 7 and 246 DF,  p-value: 0.0002081
```

Partial Models

Model 1: Internet and Urban Covariates removed, reducing selection bias of survey

Diagnosed_Diabetes ~ No_Health_Insurance + Food_Insecurity + Obesity + CIP_2021 + MHI_2021

Model 2: Including only Coefficients with p–values less than 0.1

Diagnosed_Diabetes ~ CIP_2021 + Urban

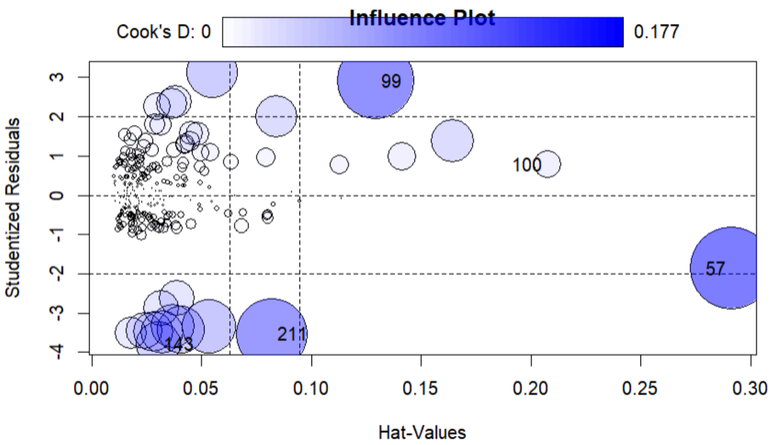
Model 3: Including Coefficients with p-values less than 0.5

Diagnosed_Diabetes ~ No_Health_Insurance + CIP_2021 + Urban

Model 2 had the best measures of AIC, BIC, and adjusted r-squared, however the low r-squared measures reflected the inability of this type of model to appropriately represent the data. This model indicated that children in poverty and whether a county was Urban or Rural had the highest correlation with the prevalence of diabetes.

Model#	df	AIC	BIC	Adj R^2
Full Model	9	189.6240	221.4600	0.08164
Model 1	7	189.0323	213.7936	0.07674
Model 2	4	180.9636	195.1129	0.09517
Model 3	5	182.3836	200.0703	0.09362

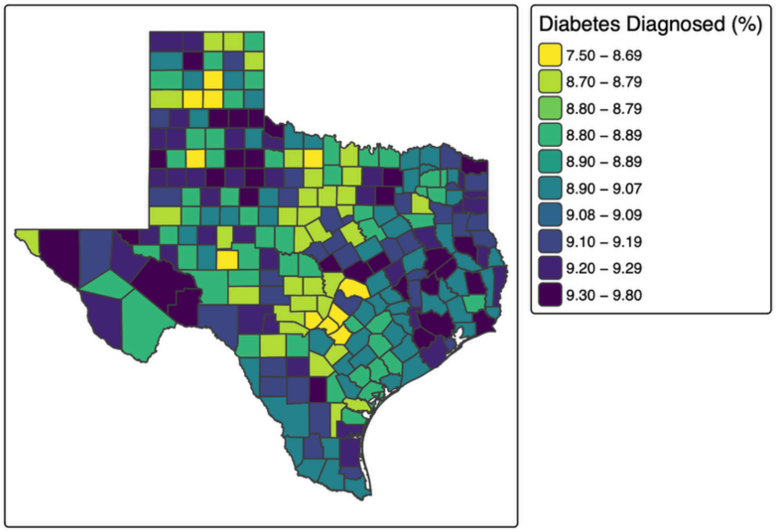
There were some outliers that appeared in all models, namely Fort Bend, Walker, Houston, Loving, and Falls Counties. Of these, all were found to be influential on the full model, and on the best model, all but Houston and Walker. While there were other influential counties, the impact of outliers would need to be considered and may be reflected in the spatial models.

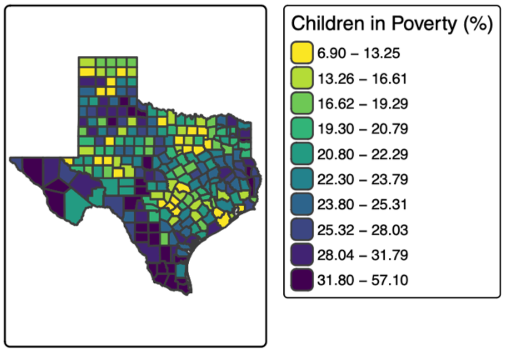


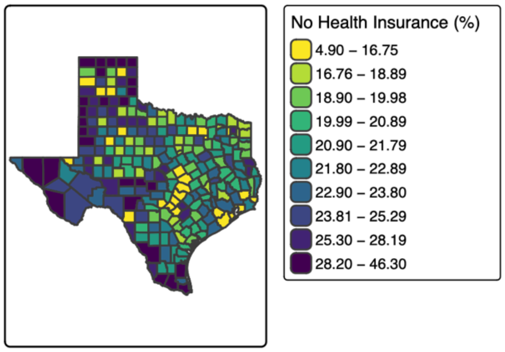
Data Modeling – Moran Test

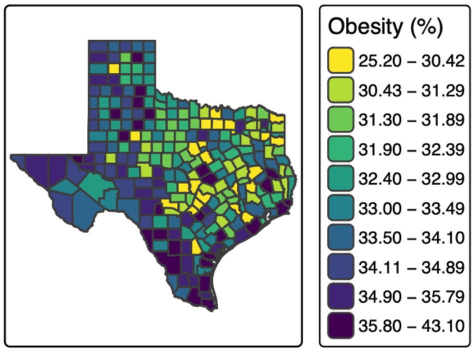
Spatial Plots

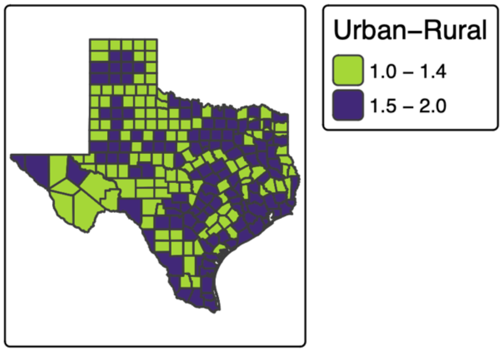
The maps show clear spatial heterogeneity in all county-level health indicators across Texas. Diabetes diagnosed (%) varies substantially by county, with clusters of higher prevalence in several southern and eastern counties and relatively lower values in many central and western counties. Children in poverty (%) and the percentage without health insurance display broadly similar patterns, suggesting that socioeconomic disadvantage is geographically concentrated and overlaps with areas of poorer diabetes outcomes. Obesity (%) is also spatially uneven and tends to be elevated in many of the same regions, reinforcing the link between obesity, poverty, and diabetes risk. The Urban–Rural map highlights that most Texas counties are classified as rural, with urban counties concentrated around major metro areas.



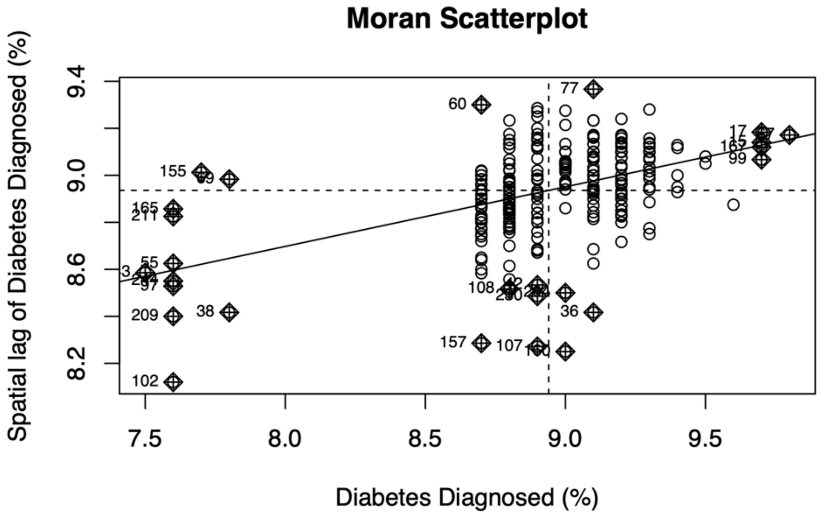








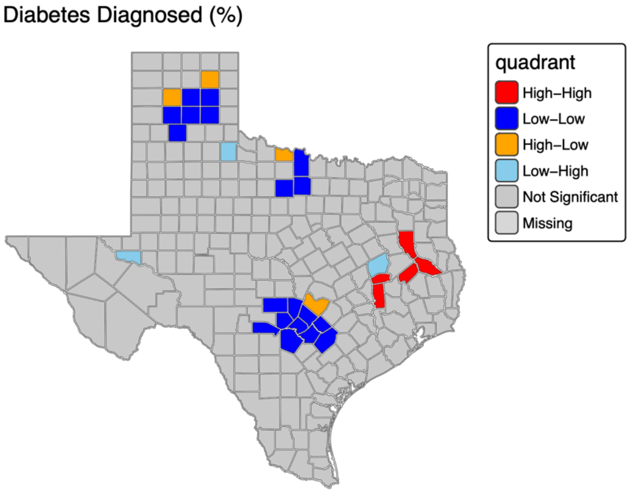
Global Moran's I Test

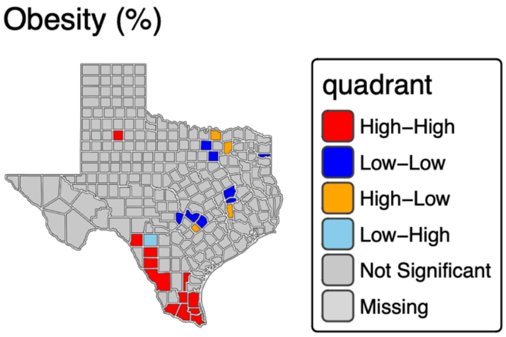


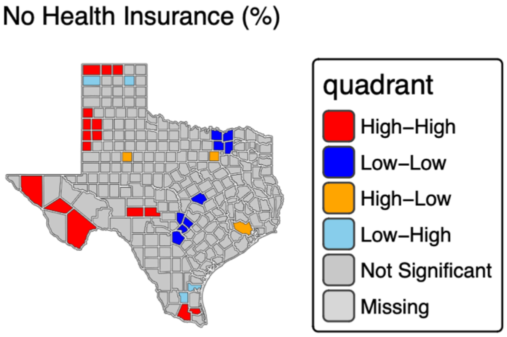
Variable	Moran I Statistic	Expected	Variance	p-value	Isolates
Diagnosed Diabetes (%)	0.2534346	-0.003952569	0.001387390	2.420643e-12	0
Children in Poverty (%)	0.4890996	-0.003952569	0.001409051	1.037715e-39	0
Obesity (%)	0.2452933	-0.003952569	0.001406232	1.499716e-11	0
No Health Insurance (%)	0.2448477	-0.003952569	0.001401327	1.502583e-11	0
Urban	0.1910158	-0.003952569	0.001429980	1.262601e-07	0

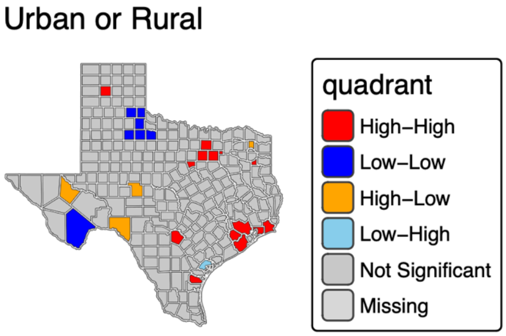
The global Moran's results show clear positive spatial autocorrelations for all variables considered. Children in poverty has the largest Moran's I (≈ 0.49, $p < 0.001$), indicating strong spatial clustering of child poverty; diabetes diagnosed (%) and obesity (%) have Moran's I around 0.25, consistent with moderate clustering, while no-health-insurance (%) and the urban rural indicator have smaller but still positive and highly significant values, suggesting weaker clustering rather than spatial randomness. The Moran scatterplots reinforce this pattern: the fitted lines all slope upward and there are no isolates, meaning each county's value tends to resemble the average of its neighbors rather than behaving independently in space.

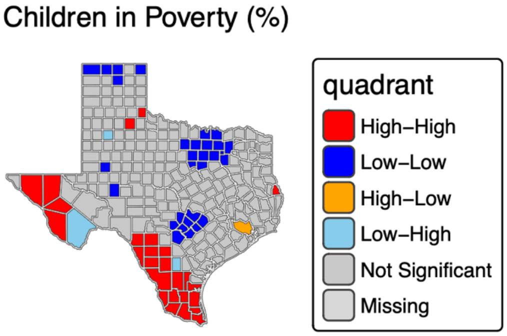
Local Moran's Test


Diabetes Diagnosed (%)


Obesity (%)


No Health Insurance (%)


Urban or Rural


Children in Poverty (%)

Variable	High-Low Counts	Low-High Counts	Interpretation
Diagnosed Diabetes (%)	9	22	Moderate Clustering
Children in Poverty (%)	27	33	Strong Clustering
Obesity (%)	16	10	Weak Clustering
No Health Insurance (%)	19	13	Weak Clustering
Urban	18	8	Weak Clustering

The local Moran analysis decomposes the global autocorrelation into county-level “hot” and “cold” spots for each variable. Across all five variables most Texas counties are grey (“Not Significant”), meaning their values are not markedly different from their neighbors once spatial dependence is considered. The interesting structure is concentrated in a relatively small set of counties that fall into the four local‑cluster types.

Diabetes Diagnosed (%): Significant High–High clusters are visible in a pocket of South-Eastern Texas border counties and High-Low and Low-Low clusters are visible in a small group in the North, indicating local hot spots of higher diabetes prevalence surrounded by similarly high counties. A few Low–Low clusters appear in parts of Central / North Texas, suggesting localized areas of comparatively lower diabetes burden.

Obesity (%): The obesity map shows several southern counties classified as High–High, forming a band of elevated obesity risk. Some Low–Low counties appear in the north, indicating local areas where both the county and its neighbors have relatively lower obesity levels.

No Health Insurance (%): Uninsured rates exhibit pronounced High–High clustering along the western border of Texas, consistent with regional pockets of elevated uninsurance. A small number of Low–Low counties occur in more central regions, indicating localized zones of better insurance coverage.

Urban or Rural: For the binary urban indicator, High–High and Low–Low patterns essentially pick out metropolitan cores versus predominantly rural regions. Some High–Low or Low–High counties at the urban–rural fringe behave as spatial outliers, where a county’s classification differs from its neighbors.

Children in Poverty (%): Child poverty shows High–High clusters in South and Western Texas, highlighting localized concentrations of high poverty. A few Low–Low clusters appear in Central and Northern parts of the state.

Overall, the local Moran maps confirm that the strongest clustering occurs for children in poverty, no health insurance, and obesity, with diabetes diagnosis and urban status showing smaller but still interpretable pockets of local clustering and a handful of spatial outliers.

Data Modeling – Kriging

The aim of this part of the project is to test a kriging algorithm for predictiveness on the 2021 diabetes spatial data set. To do this, the centroid for each county is calculated in latitude and longitude coordinates, and the diabetes dataset can be treated as point data. Figure 1 shows the diabetes prevalence by county in Texas.

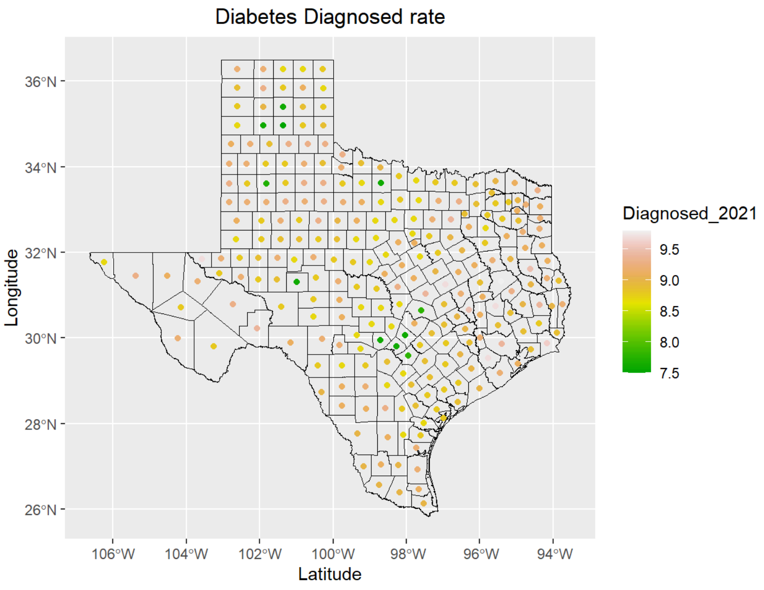


Figure 1 (Right). Occurrence of diabetes (% of population) by county in Texas in 2021.

Kriging is a method of interpolation based on a spatial stochastic process determined by a prior covariance model. It gives the best linear unbiased prediction (BLUP) for a parameter at unknown locations, given data at known locations. In this case, "best" means minimum prediction error variance, and "unbiased" means that on average for many samples, the estimated parameter will be equal to the true parameter.

The procedure for kriging in R is as follows: set up a design matrix for the trend model, consisting of 1 (for intercept), latitude, and longitude, divide the data into 80% training and 20% test points, calculate the MLE estimates of the spatial model parameters using the "likfit" function on the training data with an assumed exponential covariance model, setting up the krige calibrated model parameters with "krige.control", and then applying it to the test dataset with "krige.conv". As a QC step, the empirical semivariogram is calculated and compared to the fitted model variogram using the "variog" function.

The model parameters for the mean value (diabetes %) from MLE are" beta0 (intercept) = 10.21, beta1 (latitude) = 0.0097, beta2 (longitude) = -0.0095. The spread in latitude and longitude values is ~10, and when multiplied by their respective coefficients, show that there is minimal linear spatial trend effect (~0.1%) compared to the spread in data (~2%). The parameters for the spatial component are: partial sill = 0.12, effective range parameter (3x phi) = 0.99. This is shown as the solid line in Figure 2. The empirical points show the semivariance is approximately constant regardless of distance, and that kriging may not be an effective modeling tool for this dataset.

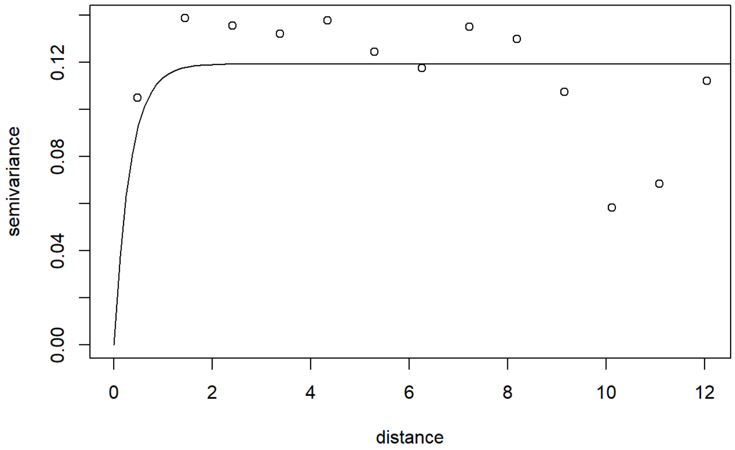


Figure 2 (Right). Comparison of the empirical semivariogram (circles) and MLE estimate for the krige model.

Plots of prediction errors at test locations are shown in Figure 3. In quantitative terms, the RMSE=0.307, r^2 = 0.287, and adj r^2 = 0.24. This shows that roughly ¼ of the dataset variance can be predicted by averaging neighboring values through ordinary kriging.

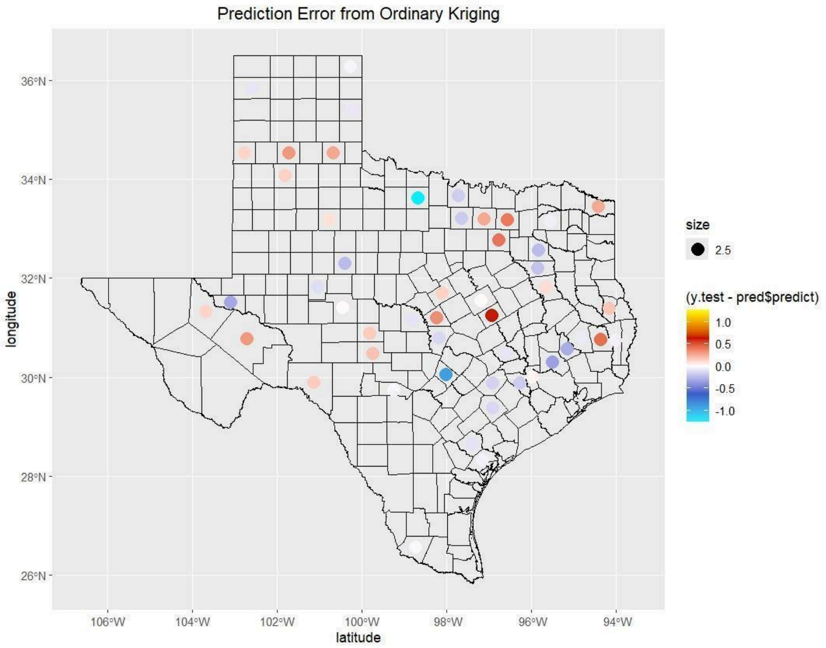


Figure 3 (Right). Prediction error from ordinary kriging for diabetes percentage of population at test locations.

Figure 4 shows the fit of individual test data points. The kriging model matches the main cloud of data around test value ~ 9.0 but does not fit the potential outlier values at test value ~ 7.6.

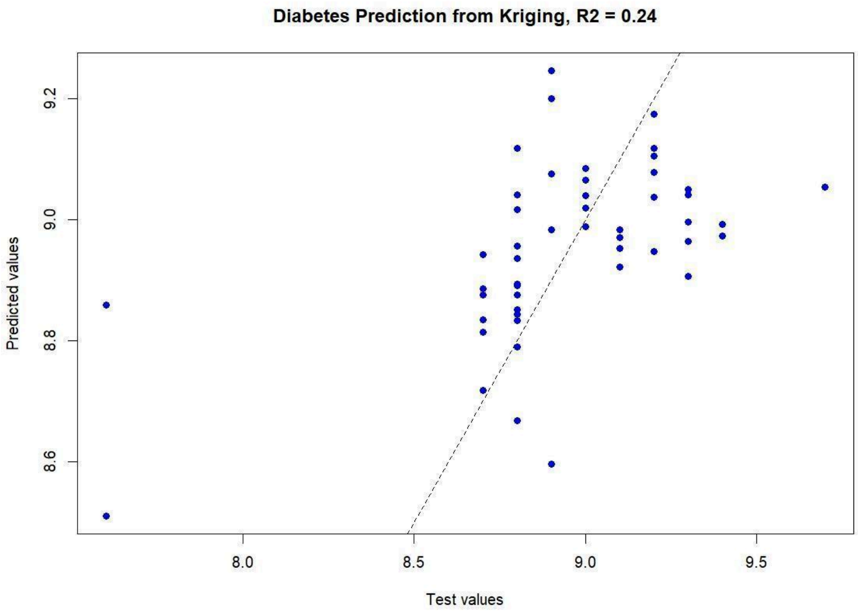


Figure 4 (Right). Diabetes prediction from ordinary kriging at test points (blue). Dashed black line is 1:1 prediction.

Data Modeling – Spatial Temporal Gaussian Process Model

The next step in model complexity is to use a spatial temporal Gaussian process model to predict the occurrence of diabetes through time. This enables both spatial weighting of surrounding data and time-varying regression equations based on likely predictive factors at the exact location (covariates). For the diabetes dataset, years 2017-2021 were considered for the analysis, as 2017 was the earliest year available. This dataset is plotted in Figure 5. There is generally a slight increase in diabetes over time, with the exception of the final year (2021). Note that because of the limited reporting precision, many data points plot on top of each other, and the average by year is shown for clarity.

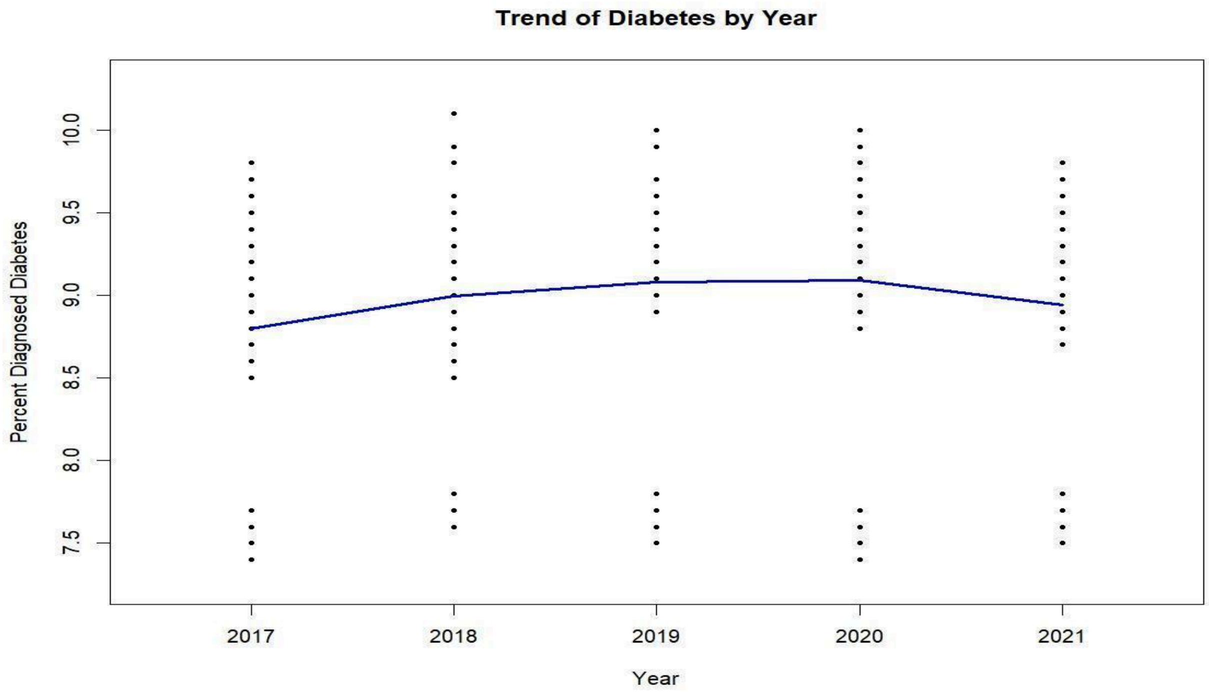


Figure 5. Percent of population diagnosed with diabetes by year, solid blue line is the average by county.

The R function ‘spDynLM’ is used to fit Gaussian univariate Bayesian dynamic space-time regression models where space is viewed as continuous and time intervals are discrete. Documentation for spDynLM is located at:
https://www.rdocumentation.org/packages/spBayes/versions/0.4-8/topics/spDynLM. This function fits a Gaussian spatial-temporal regression model and uses the nearest neighbor approximation for Markov Chain Monte Carlo (MCMC) simulation. The ‘apply’ function makes predictions at locations, based on the calibrated model. The model is specified in form:

$$y_t(s) = x_t(s) + \ u_t(s) + \ \epsilon_t(s), \ t \ = 1, 2, ...N_t$$

$$\epsilon_t(s) \sim N(0, \tau_t^2)$$

$$\beta_t = \beta_{t-1} + \eta_t; \ \eta_t \sim N(0, \Sigma_\eta)$$

$$u_t(s) = u_{t-1}(s) + w_t(s); \ w_t(s) \sim \mathrm{GP}(0, C_t\big(., \theta_t\big)$$

$$C_t\big(., \theta_t\big) = \sigma^2 \rho\big(s_{1,} s_{2,} \phi_t\big)$$

Where $y_t(s)$ is the prediction at time t for the spatially varying process,

$x_t(s)$ is the design matrix at time t, with coefficients β_t, that depend on the those of the previous time step plus additional variation η_t, where

η_t is normally distributed with mean=0, and specified covariance matrix Σ_η

$u_t(s)$ is the spatially varying process for time t, which depends on that of the previous time step with the addition of an additional spatially varying process for the current time step, $w_t(s)$,

$w_t(s)$ is the latent spatial process that accounts for spatial autocorrelation, specified with mean value of zero and covariance function $C_t\left(., \theta_t\right)$,

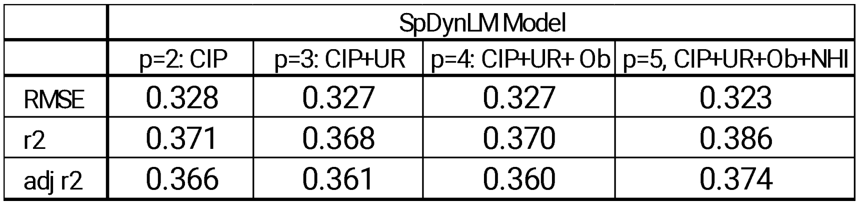
$C_t\left(., \theta_t\right)$ is specified by spatial variance σ^2, spatial correlation function $\rho(., \phi_t)$,

ϕ_t is the range parameter that controls decay of correlation with distance.

$\epsilon_t(s)$ is spatially-independent residual noise at time t, specified by a normal distribution with mean=0, and variance = τ_t^2.

A variety of covariates were tested for the design matrix. From the ordinary least squares regression, the only significant predictor at p-value = 0.05 is percentage of children in poverty (CIP). The first case for p=2 uses intercept plus CIP. Additional cases were tested with the optimal covariates for p=3 determined to be CIP + UR (Urban/Rural). For p=4, the optimal covariates are CIP + UR + Ob (Obesity). Finally, for p=5, CIP+UR+Ob+NHI (no health insurance) is used, as these were the only variables with measurable effect (even if not strictly significant). The dataset was split into ~80% training values and 20% test values, with random seed fixed for equivalency between the different trials. To overcome starting parameter effects, the first quarter of realizations were discarded. For the simulation, starting values, tuning parameters, prior distributions for spatial parameters spatial range (phi), spatial process variance (σ^2), and spatially independent noise (τ^2) were specified through trial-and-error adjustment with multiple trials of 4000 MCMC samples.

To determine the optimal number of covariates, error measure RMSE, r^2, and adj r^2 were compared and are listed in table 1. The difference in RMSE and adj r^2 is minimal going from predictor CIP to 4 predictors CIP+UR+Ob+NHI. The adj r^2 of 0.37 is significantly higher than the 0.24 from kriging, showing that inclusion of the CIP covariate improves the predictability of diabetes. However, additional covariates add negligible predictability.



	SpDynLM Model			
	p=2: CIP	p=3: CIP+UR	p=4: CIP+UR+ Ob	p=5, CIP+UR+Ob+NHI
RMSE	0.328	0.327	0.327	0.323
r2	0.371	0.368	0.370	0.386
adj r2	0.366	0.361	0.360	0.374

Table 1. Comparison of prediction error measures from spatial temporal regression modeling with different covariates.

As a QC of the process, the spatial model parameters as a function of time for p = 2 are plotted in Figure 6. For a p-value = 0.05, the differences in spatial model parameters through time are not significant. Additional investigation shows that the difference in spatial variance for 2017 is significant at p-value = 0.1.

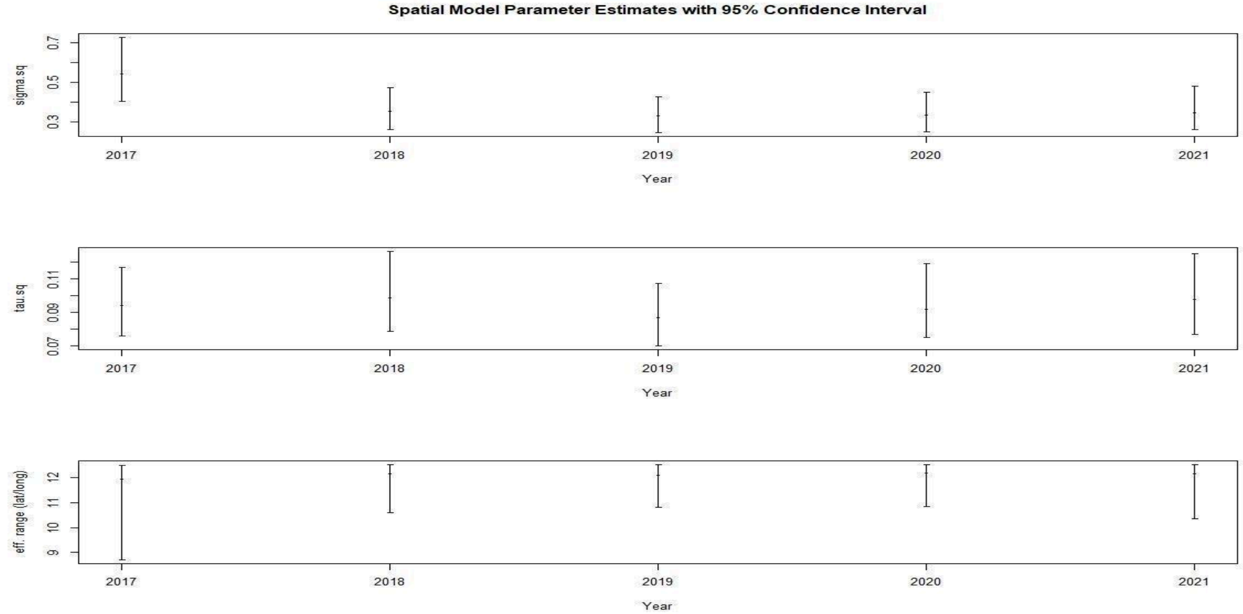


Figure 6. Spatial model variance (sigma.sq), residual noise (tau.sq) and effective range for spatial temporal model with p = 2 (CIP).

Regression model coefficients (beta) are shown in Figure 7. With a p-value of 0.05, changes in intercept of CIP coefficient are not significant from years 2017 to 2021. That is, there exist possible constant values of the coefficients that are contained within the 95% confidence interval for all years. The mean value of beta.1 (CIP) is consistently above 0 for all years 0 is outside of the 95% confidence interval for all years except 2021, showing that CIP is generally significant as a predictor.

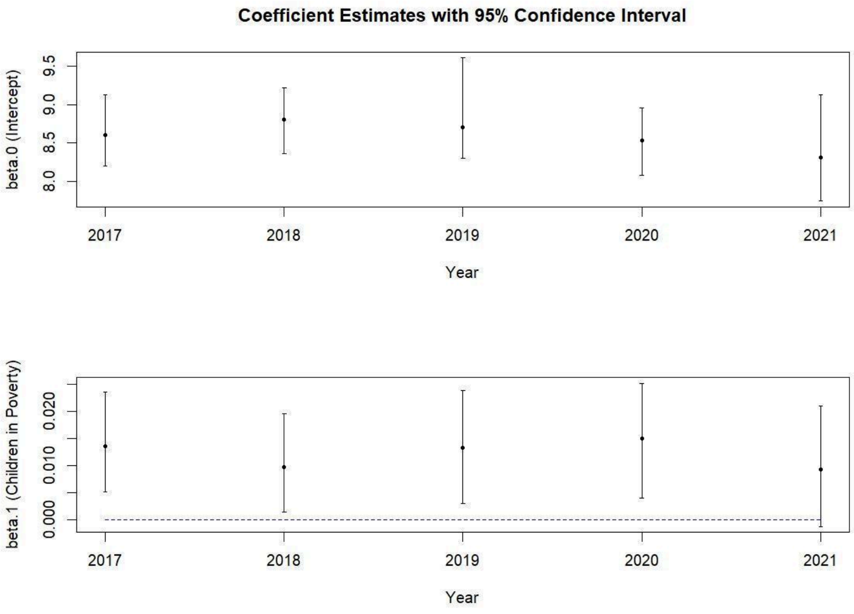


Figure 7. Model coefficients (beta) for p=2 (CIP). Dashed line is value of 0.

As a QC, Figures 8 and 9 show the coefficients for the model with p=5 (CIP+UR+Ob+NHI). Adding additional covariates from the p=2(CIP) model enlarges the confidence interval for beta.1(CIP). The beta.1(CIP) confidence intervals contain the value zero for years 2018, 2019, 2021 meaning that it cannot be ruled out that CIP has no effect on diabetes occurrence for these years, with a p-value of 0.05.

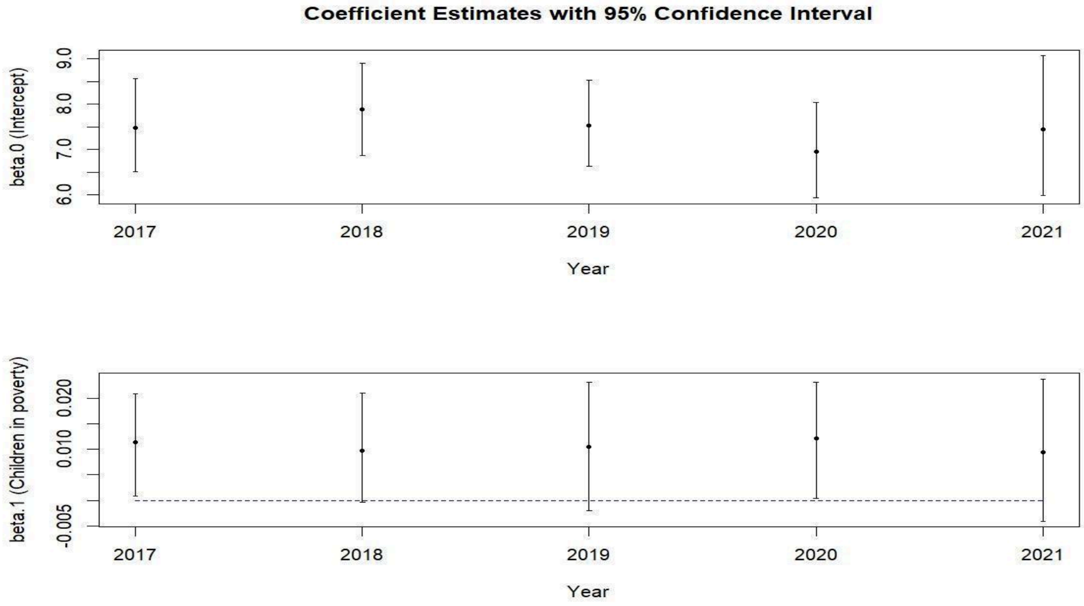


Figure 8. Model coefficients (beta) for p=5 (CIP+UR+Ob+NHI). Dashed line is value of 0.

The additional beta coefficients for the p = 5 (CIP+UR+Ob+NHI) model are not significantly different than zero at p-value = 0.05 (Figure 9), with the exception of beta.4(NHI) in 2020. This is consistent with these additional predictors adding minimally to the adjusted r^2 measure (see Table 1).

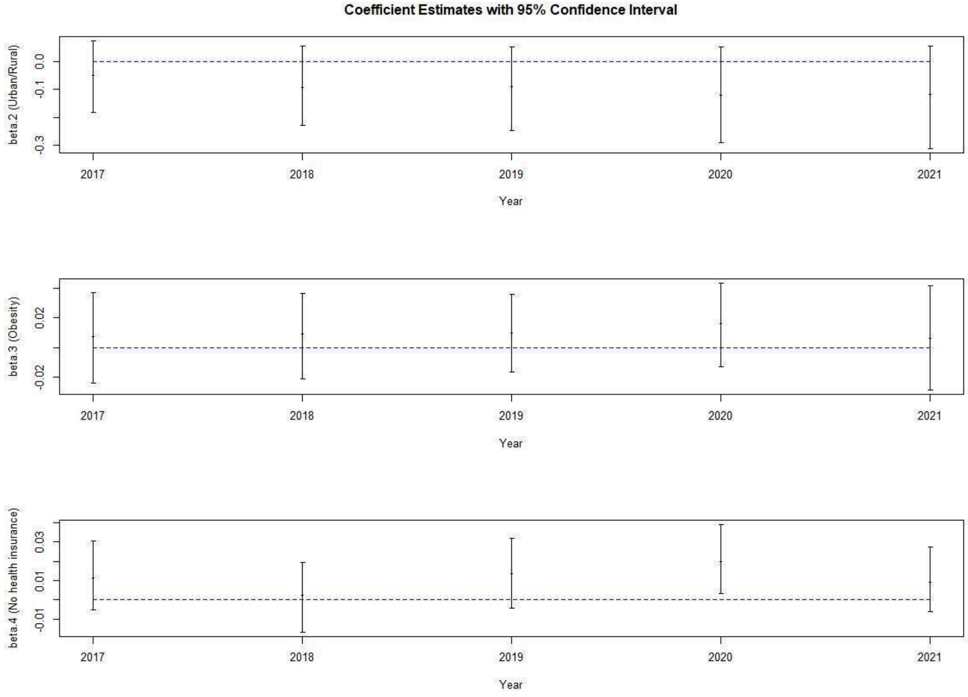


Figure 9 (Left). Model coefficients (beta) for p=5 (CIP+UR+Ob+NHI). Dashed line is value of 0.

Figure 10 shows the fit of spatial-temporal Gaussian regression model for the training and test data for all years. The error scatter in the test data is generally wider than for the training data. Like the kriging model, the spatial Gaussian model does not do well at predicting the anomalously low values of diabetes percentage near 7.5.

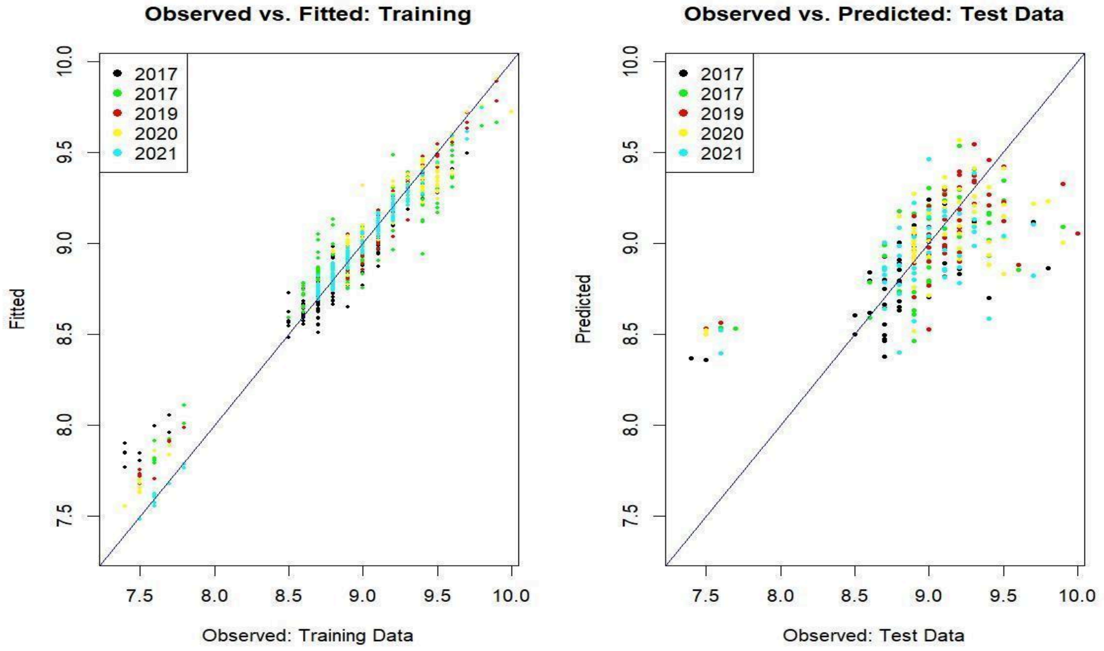


Figure 10. Fit of training data and predictions for test data for all years of diabetes: percentage diagnosed.

Figure 11 shows the fit of spatial-temporal Gaussian regression model for the test data in 2021 with 95% prediction intervals. The 95% prediction interval for each county data point is approximately the original range of the entire data set ~7.5 to ~10. This is consistent with the model having modest adj r^2 of 0.37. leaving most of the variance unexplained by the model and leading to wide prediction intervals. Figure 12 shows the prediction error for year 2021 at the test locations.

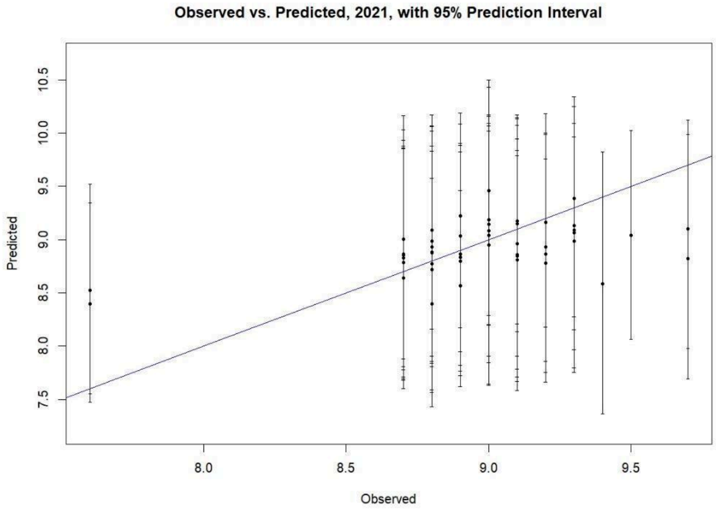


Figure 11 (Left). Fit of test data from spatial-temporal model for year 2021, with 95% prediction intervals. Note due to limited reporting precision, many data points have similar x (observed) value. 1:1 line superimposed.

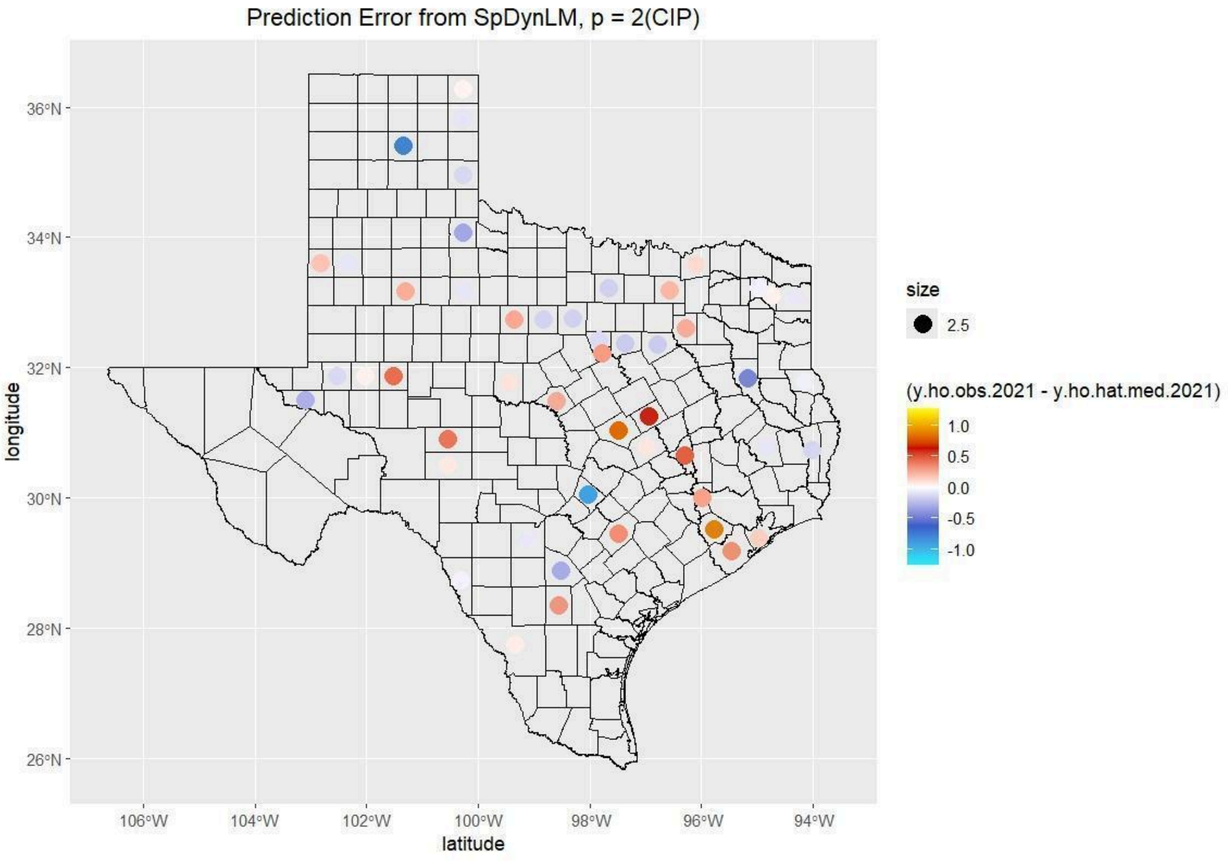


Figure 12. Prediction error from spatial temporal Gaussian regression for diabetes percentage of population at test locations in 2021.

Discussion – Kriging and Spatial Temporal Gaussian Process Model

Ordinary kriging of the 2021 diabetes dataset yields low predictivity, with adjusted r^2 of ~0.24. The spatial temporal Gaussian regression model improves on this by including additional predictive data. The best model is for p=2: an intercept and covariate 'Children in poverty.' This model increases the adjusted r^2 to 0.37. Including additional covariates such as 'Urban/Rural', 'Obesity', 'No health insurance', has minimal effect. There is little 'significant' change in spatial model parameters or model coefficients with time from 2017 to 2021, for p-value = 0.05. Why is this? The model already has large uncertainty. Perhaps there are unmeasured covariates that could explain more of the diabetes diagnosis data, or perhaps there is large unpredictable variance. It may also be that 5 years is too short of a time interval to detect change with statistical significance. It may also be that the covariates such as obesity, need a larger time horizon than 1 year to take effect. If diabetes is a chronic rather than acute condition, with newly diagnosed patients adding to a much larger long-term-diagnosed population, the covariates may be better at predicting newly diagnosed cases. In addition, there is mobility within counties over time, so that the prior health history of the population changes. To test this, data (including newly diagnosed patients by year) would be needed over decades, or better yet, on a large population of individuals tracked over their lifetime. An autoregressive model could be tested to see the time lag effects of health, lifestyle, and economic factors.

Data Modeling – Geographically Weighted Regression

Geographically Weighted Regression (GWR) analyzes how the effects of independent variables (e.g., obesity prevalence, Food Environment Index, etc.) on a dependent variable (diagnosed diabetes) change across locations. While a multiple linear model has the advantage of showing the relationship between the dependent variable and the independent variables at a glance, this approach allows a deeper analysis of spatially varying effects.

Key Findings and Discussion

Spatial Patterns of GWR Coefficients

The coefficient maps visualize, for each factor, how strongly it affects diabetes across space (Figure 13). Depending on the inherent characteristics of each factor, the magnitude of its effect on diabetes varies geographically. For example, the Chronic Disease Composite Index (CIP score) shows a relatively strong influence on diabetes in southeastern Texas, including the Bryan and Austin areas. In contrast, the percentage of adults without health insurance does not have a strong effect in this region. The Food Environment Index has a relatively large impact in metropolitan districts classified by TxDOT (Dallas, Fort Worth, Houston, and San Antonio), but it does not show a meaningful influence on diabetes in other regions. These results suggest that, even within the same TxDOT district, the priority of factors to focus on can differ by location.

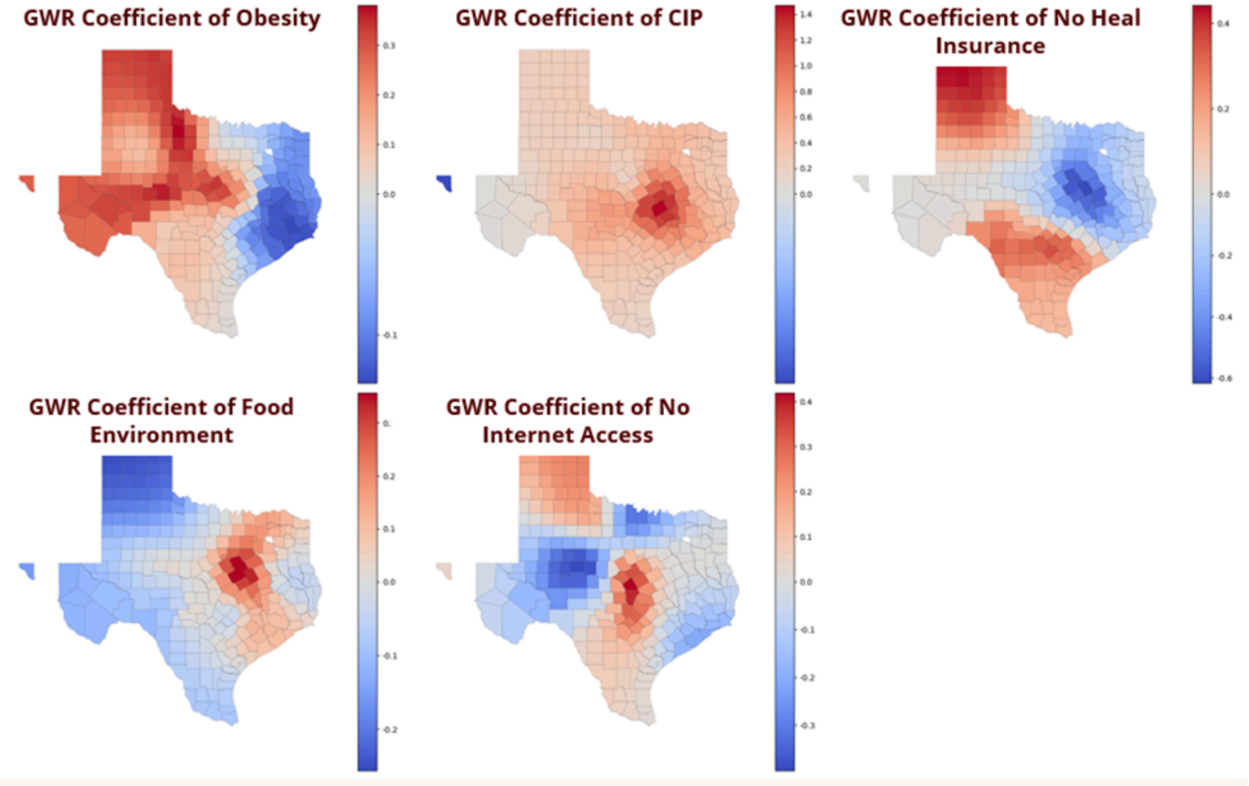

GWR Coefficient of Obesity
0.3
0.2
0.1
0.0
-0.1
GWR Coefficient of CIP
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
GWR Coefficient of No Heal Insurance
0.4
0.2
0.0
-0.2
-0.4
-0.6
GWR Coefficient of Food Environment
0.
0.2
0.1
0.0
-0.1
-0.2
GWR Coefficient of No Internet Access
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3

Figure 13. Spatial variation of GWR coefficient for diabetes risk factors.

Spatial Local R^2

The GWR-based local R^2 explains a larger share of diabetes variation in central and southern areas, where many people are concentrated, including metropolitan districts (Figure 14). Since most of Texas's population lives in these areas, this result shows the usefulness of the GWR model based on the selected factors. In contrast, in the northern and northwestern regions, which are less populated and have relatively harsh environmental conditions, the explanatory power of the GWR model based on the considered factors is lower (Figure 15). Therefore, to analyze diabetes in these regions, additional factors (e.g., unmeasured behavioral risks) need to be considered.

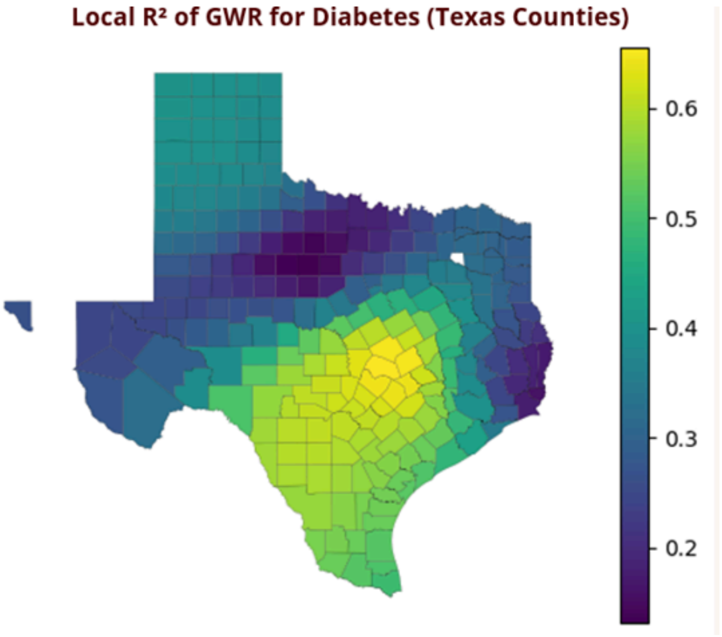


Figure 14. Spatial distribution of local R^2 from GWR (Texas)

Furthermore, when counties are classified as urban or rural, the average local R^2 for urban counties shows a slightly higher tendency. However, overall, no meaningful difference is found between the two groups.

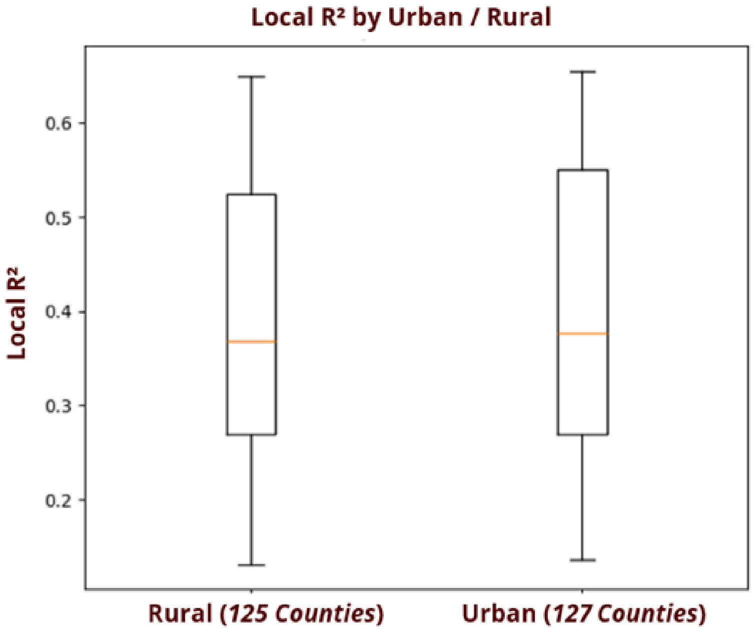


Figure 15. Comparison of local R^2 between rural and urban counties.

Spatial Changes in Dominant Risk Factors Over Time

A. The southern region shows no consistently dominant risk factor

In the counties of the southern region over the three years analyzed, no single factor is identified as a dominant factor (Figure 16). This does not mean that there is no meaningful relationship between diabetes and the factors used in the study. Rather than one factor having a clearly strong effect on diabetes, multiple factors are likely to interact with each other to influence diabetes. Additional analysis is needed to understand how these factors are related and how they jointly affect diabetes. This result may also indicate that some important factors influencing diabetes were not included. Therefore, in this region, it is necessary to consider additional social, occupational, or clinical factors beyond those used in the GWR analysis.

B. Obesity has become a stronger driver in western Texas

In the western region, obesity begins to be identified as the dominant factor in the most recent year (2020). This suggests that special attention to obesity is needed to prevent diabetes in this region, and that district-level obesity improvement policies may be required. Further in-depth study is needed to examine whether the pandemic contributed to obesity emerging as a dominant factor during this period.

C. No-internet access gains importance in metropolitan corridors after 2019–2020

In metropolitan areas, the no-internet-access factor begins to show a particularly strong association with diabetes in 2020. This does not mean that metropolitan areas have many places with no internet access overall. Rather, it should be understood that within cities, there are certain areas where internet connectivity is especially low, and diabetes prevalence is much higher in those areas. Similar to point B, further analysis is needed to examine how the global pandemic may have influenced this result.

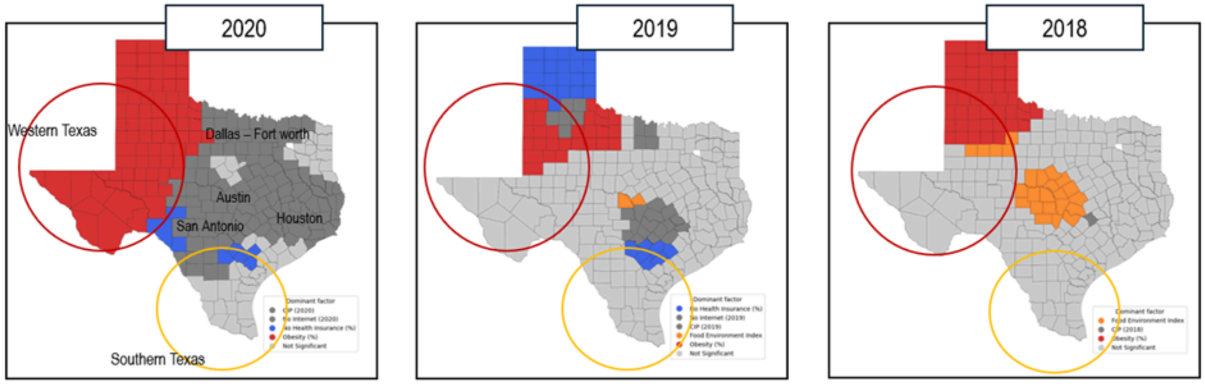


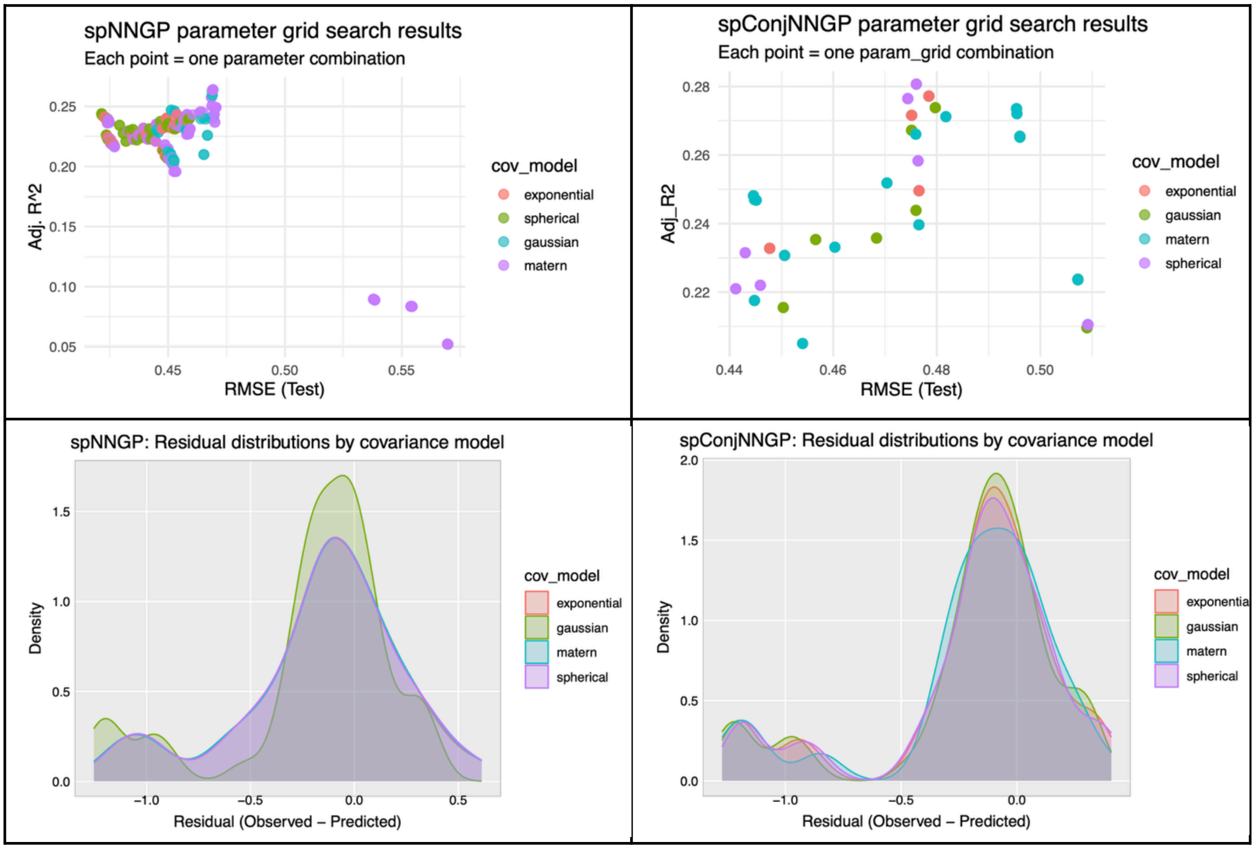
Figure 16. Changes in Dominant Diabetes Risk Factors (2018–2020)

Data Modeling – Nearest Neighbor Models

Here a hyper-parameter grid is constructed to explore alternative NNGP specifications. Four covariance families (exponential, spherical, Gaussian, and Matérn) are considered. For each family, combinations of prior settings for the variance components (σ^2, τ^2) and range parameter (ϕ), Metropolis tuning scales, starting values for (σ^2), (τ^2), (ϕ), and several neighbor sizes (10, 15, 25) are formed using expand.grid. For the Matérn model an additional smoothness parameter (ν) is included. Overall, the grid contains 480 distinct parameter configurations (for spNNGP) and 240 distinct parameter configurations (for spConjNNGP), which were used to fit and compare a large set of NNGP models in the subsequent multi-threaded search.

Spatial Nearest Neighbor Gaussian Process Model (spNNGP)

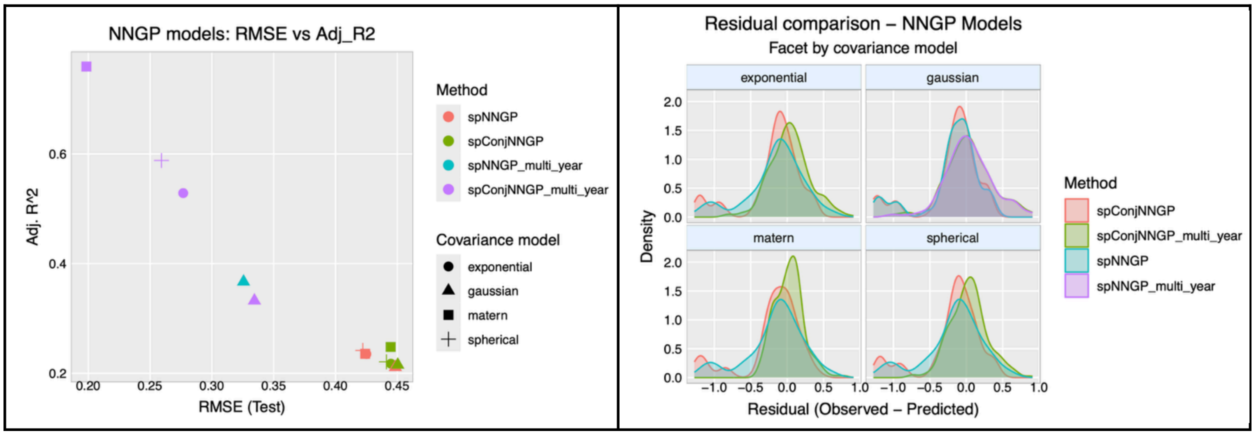
We fit spNNGP nearest-neighbor Gaussian process regression models for Texas county diabetes prevalence using obesity, no health insurance, children in poverty, and urban–rural status, comparing exponential, spherical, Gaussian, and Matérn covariance structures. The grid search shows most parameter combinations landing in a tight performance region (test RMSE ≈ 0.42–0.45; adjusted (R^2) ≈ 0.21–0.25), suggesting stable predictive behavior across priors/starting values/neighbor sizes. Among the best-per-covariance refits, spherical and Matérn perform slightly better (lowest RMSE ≈ 0.422–0.424; highest adjusted (R^2) ≈ 0.24), with exponential close behind and Gaussian a bit weaker. Residual density curves are centered near zero and look very similar across covariance choices, supporting a well-calibrated fit where remaining errors are small and largely random.



Spatial Conjugate Nearest Neighbor Gaussian Process Model (spConjNNGP)

Using the same response, split, and covariates, spConjNNGP provides a conjugate (computationally efficient) nearest-neighbor alternative evaluated over covariance families, priors, and neighbor sizes (10 vs 25). The grid-search scatter shows moderate, consistent predictive performance overall (test RMSE clustered ≈ 0.44–0.45; adjusted (R^2) ≈ 0.22–0.25). The Matérn model with 25 neighbors attains the highest adjusted (R^2) (≈ 0.25) at RMSE comparable to the other models, while spherical/exponential can achieve slightly lower RMSE with only marginal differences in explained variation. Residual distributions across covariance families are nearly overlapping and centered near zero, indicating no major lack-of-fit driven by the covariance choice and confirming broadly comparable performance across the best spConjNNGP specifications.

Temporal NNGP



Across covariance families, the single-year spNNGP and spConjNNGP best models show very similar predictive accuracy, clustering around test RMSE ≈ 0.42–0.45 and adjusted (R^2) ≈ 0.21–0.25; spConjNNGP tends to have a slightly higher RMSE than spNNGP, but adjusted (R^2) differences are small and inconsistent. The clear improvement comes from incorporating time via the multi-year specifications: multi-year NNGP models shift sharply toward lower RMSE (≈ 0.20–0.33) and higher adjusted (R^2) (up to ~0.75), with Matérn and spherical versions standing out. Residual density comparisons reinforce this, as all methods remain centered near zero (little bias) but the multi-year fits have noticeably tighter residual distributions, indicating reduced error variability and more stable predictions overall.

Discussion – Nearest Neighbor Models

This analysis shows that diagnosed diabetes prevalence in Texas is strongly structured across space and time and aligns closely with county-level risk factors—higher obesity, higher child poverty, higher uninsured rates, and more rural status. Global Moran’s I confirms significant positive spatial autocorrelation in diabetes and key covariates, while Local Moran’s I pinpoints coherent hotspot and coldspot regions with relatively few spatial outliers, reinforcing that diabetes burden is geographically clustered rather than randomly distributed.

Single-year spatial modeling with spNNGP and spConjNNGP yields stable but moderate predictive performance across covariance choices (RMSE ≈ 0.42–0.45; adjusted (R^2) ≈ 0.21–0.25), with diagnostics indicating broadly unbiased residuals but some remaining difficulty at the extremes of prevalence. The conjugate spConjNNGP fits are competitive but do not meaningfully exceed the MCMC-based spNNGP in the single-year setting. The major improvement comes from the temporal extension: multi-year NNGP models substantially reduce prediction error (best RMSE ≈ 0.20–0.33) and sharply increase explained variation (adjusted (R^2) up to ~0.75), with residuals remaining centered near zero and showing more compact variability across years. Overall, the results support using multi-year NNGP approaches—particularly Matérn or spherical specifications—to best capture the combined spatial dependence and year-to-year structure in Texas diabetes prevalence while maintaining well-behaved residuals and consistent performance over time.

Final Conclusions

In summary, of the models tested, the most appropriate model to account for the data was a multi-year spConjNNGP with a Matérn covariance, as it had not only the highest adjusted R^2 and low standard error, but it also accounted for spatial and temporal variability. We determined that this spatial dependence was present and necessary to account for with the Moran's tests.

The covariates with the highest impact on diabetes prevalence changed over time and were clustered in spatial groupings. These high-impact factors were mainly Children in Poverty and Obesity in the most recent years. While obesity is a condition that occurs alongside diabetes, it is notable that the socioeconomic factor of poverty had such a high correlation. It also did not appear to be solely because of lack of access to health care, as the uninsured rates did not have as high of a correlation, though it was not entirely insignificant.

Model	Adjusted R^2 (Of Best Model)
OLS	0.095
Kriging	0.24
STGP	0.37
NNGP	0.759

Diabetes prevalence itself did not significantly increase or decrease over five years according to the data, but it itself was spatially distributed, with a higher prevalence in rural environments, which is counter to the general expectation. In future study, it would be useful to utilize a data set that has a larger time frame to see if the trends follow a similar pattern, or if these factors are largely a more recent occurrence.

References

1. Analysis - United States Diabetes Surveillance System. https://gis.cdc.gov/grasp/diabetes/diabetesatlas-analysis.html#. Accessed November 9, 2025.

2. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System www.cdc.gov/brfss. Accessed September 2025.

3. US Census Bureau. Population Estimates Program. Population and Housing Unit Estimates. https://www.census.gov/popest/. Accessed September 2025.

4. Rao JNK. *Small Area Estimation*. Hoboken, New Jersey: John Wiley & Sons, Inc; 2003.

5. Cadwell, B. L., Thompson, T. J., Boyle, J. P. and Barker, L. E. (2010). Bayesian small area estimates of diabetes prevalence by US county, *J Data Sci.* 2005;8(1): 173-188

6. Barker LE, Thompson TJ, Kirtland KA, Boyle JP, Geiss LS, McCauley MM, Albright AL. Bayesian small area estimates of diabetes incidence by United States county, 2009. *J Data Sci.* 2013;11:249-269.

7. 2013 NCHS Urban-Rural Classification Scheme for Counties. https://www.cdc.gov/nchs/data_access/urban_rural.htm. Accessed September 2025

8. Data & documentation. County Health Rankings & Roadmaps. https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation.

9. US Census Bureau. 2008 - 2023 Small Area Health Insurance Estimates (SAHIE) using the American Community Survey (ACS). Census.gov. Published November 25, 2025. https://www.census.gov/data/datasets/time-series/demo/sahie/estimates-acs.html