Predicting Chase Bank Money Deposits in Ohio State from 2010 to 2016 using Spatio-Temporal kriging

Finnegan Nguyen

16 December 2019

Contents

References

1. Introduction

1.1. Chase Bank Money Deposits and Question of Interest

Bank deposits are money place into banking institutions such as savings accounts, checking accounts and money market accounts by the account holder (customer). People can deposit in banks in the form of cash, paper checks and electronic transfers. Bank deposits are important because they stimulate economic growth. Banking institutions lend some of the deposited money to many people to help them buy houses, cars, or even start businesses. As of 2019, JPMorgan Chase has the most deposits in the US with $1.139 trillion (Banks, 2019). However, Chase still not dominate the deposits market in Ohio. One way to increase the amount of deposits is to open more branches so people in every area can access the bank easier.

The aim of this report is to use universal kriging method by choosing adequate spatio-temporal variogram to predict money deposits at unobserved locations in Ohio state for each year from 2010 to 2016 from data observed at 264 known locations.

1.2. The Data

The data used in this report relates to the amount of money deposits in each Chase branches in Ohio state. Theses yearly data is obtained from the Summary of Deposits originated from https://www.fdic.gov. The Summary of Deposits is the annual survey of each branch office deposits of all institutions that is insured by Federal Deposit Insurance Corporation (FDIC).

The data set consists of one variable that is yearly money deposits (deposits) in thousands of dollars ($000) at 264 Chase branches in Ohio state (between 38^{o}N-42^{o}N and 81^{o}W-85^{o}W), recorded between the years 2010 and 2016. These data are changed with time (yearly) and irregular in space. The data are complete with no missing value but the branches themselves are obviously not located everywhere in Ohio. Therefore, if Chase has knowledge about the amount of money deposits in the whole area they can consider where to open the next locations.

Despite the shift to online banking, people still find the need to go to bank braches. On average, about 4 in 5 Americans visit a bank at least one a year. Especially, if one wants to deposit a large amount of money he or she feels more secure to get it done by a bank's agents. In fact, people with higher income would likely to visit banks more. When Chase entered the market in Ohio in 2004, they opened their branches in big city first then eventually opened more in other area. Ohio is an interesting place to explore the spatial structure of bank locations because Chase is currently trying to gain more customers here. Figure 1 shows that bank locations are not randomly distributed. Most banks are located in the Central, North East and South West area of Ohio.

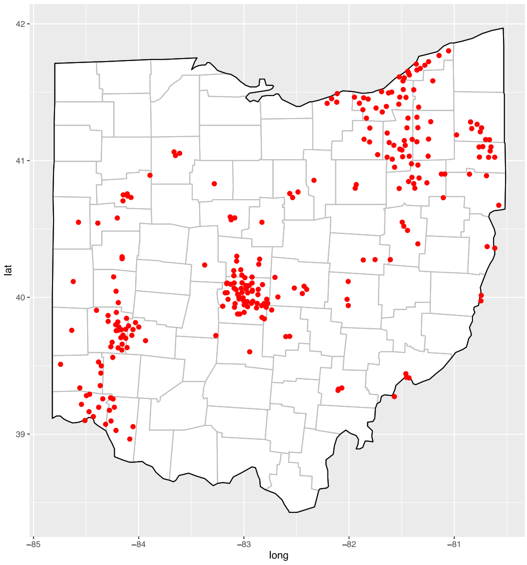


Figure 1. Chase bank locations
in Ohio state as of 2016.

1.3. Visualization of Spatio-Temporal Data

The deposits range from 1 million to 827 millions. By taking the logarithm of the data, the amount of deposits are more interpretable.

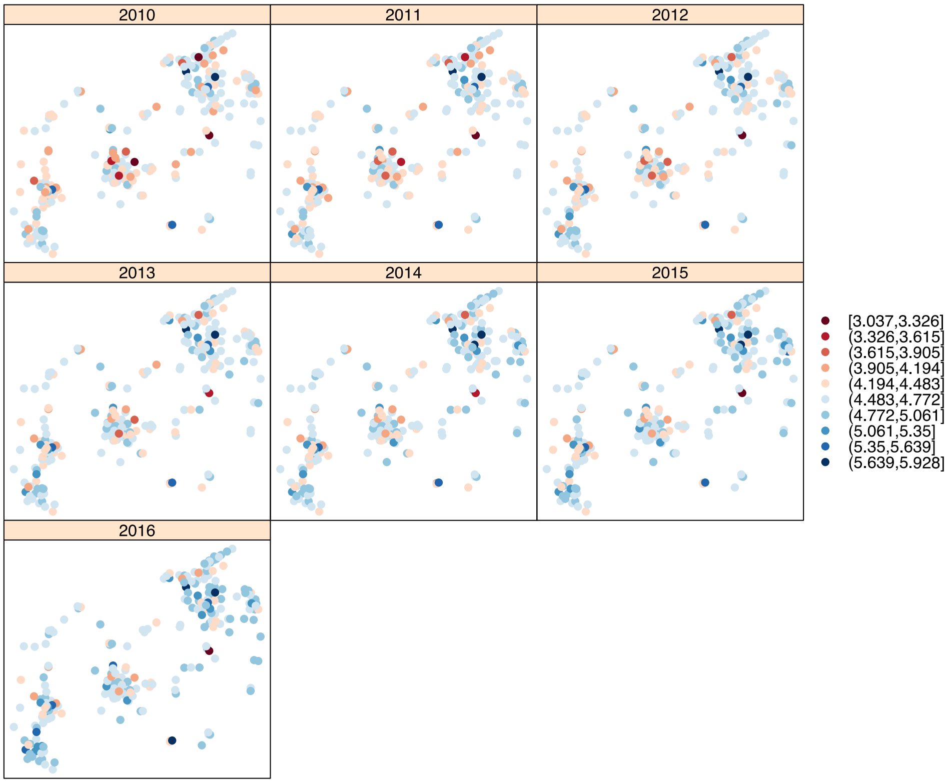


Figure 2. Money deposits (in $000) from the Chase bank
data set for each year from 2010 to 2016, plotted on a log

In Figure 2, The red dots are less money and the blue dots are more money. Deposits are increasing through time. At each bank location, the amount of deposits is relating to the bank around it in the same area. There is a relationship between the observational locations and their amount of deposits. Therefore, there exist a spatial correlation among the data.

2. Methods

2.1. Spatial Correlation, the Pooled Variogram

In this project, the sample consists of 7 time instances (7 years). These instances are SpatialPointDataFrame and are added a time index to them. After, these points are bonded in a single SpatialPointDataFrame which has a time index ti. The pooled variogram in Firgure 2 has a range about 10 km and a sill about 0.65. The fitted model is Matern, M. Stein's parameterization with the partial sill of 0.6461874, the range of 2.182146 and the kappa of 0.5. In Figure 3, the fit is rather poor. In this case, the exploration of a spatio-temporal variogram is needed to construct a better fitted model.

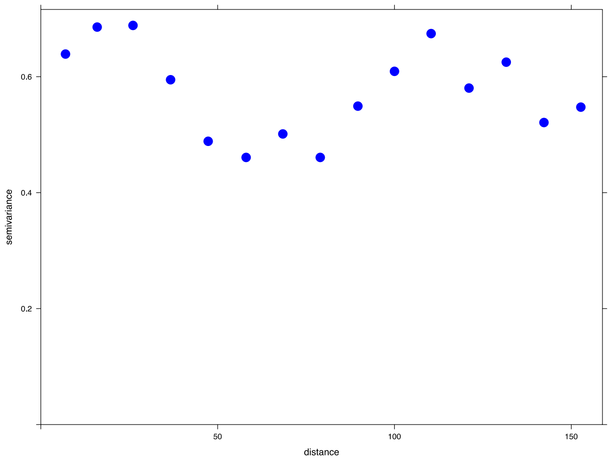


Figure 2. The pool variogram, average over 7 chosen time steps.

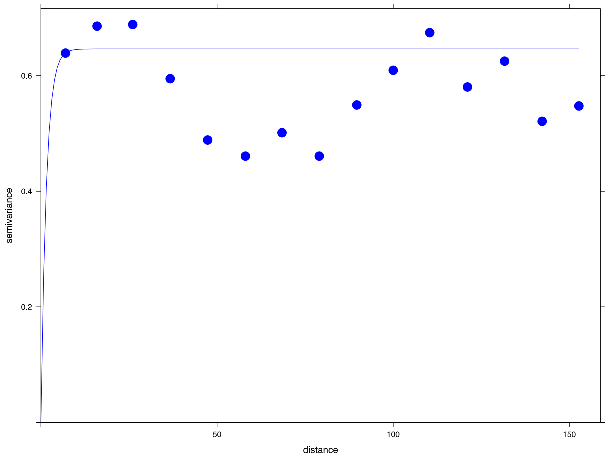


Figure 3. The fitted pooled variogram, average over 7 chosen

2.2. Empirical Spatial-Temporal Variogram

Assuming Z = $\{ Z(s_1; t_1), \ldots, Z(s_n; t_n)\}$ is the spatio-temporal data sets. The spatio-temporal variogram is defined as

$$\gamma(s_i, s_k; t_j, t_l) = \frac{1}{2} Var(Z\big(s_i; t_j\big) - Z(s_k; t_l)$$

The empirical variogram is used as a first estimate of the theoretical variogram needed for spatial interpolation by kriging. In the case where the covariance depends only on displacements in spaces and differences in time, the empirical variogram can be written as

$$\hat{\gamma}(\boldsymbol{h}; u) = \hat{C}(\boldsymbol{0}; 0) \; - \; \hat{C}(\boldsymbol{h}; u)$$

$\boldsymbol{h}$ = $s_k - s_i$ is spatial lag

$u = t_l - t_j$ is a temporal lag

The gamma ranges from 0.00 to 0.16. The lower is the gamma, the higher is the correlation. In Figure 4, the pink and purple blocks are 0.09 and 0.06 respectively. These blocks between 60 km and 80 km and from 0 days to 1500 days (5 years) have a high correlation.

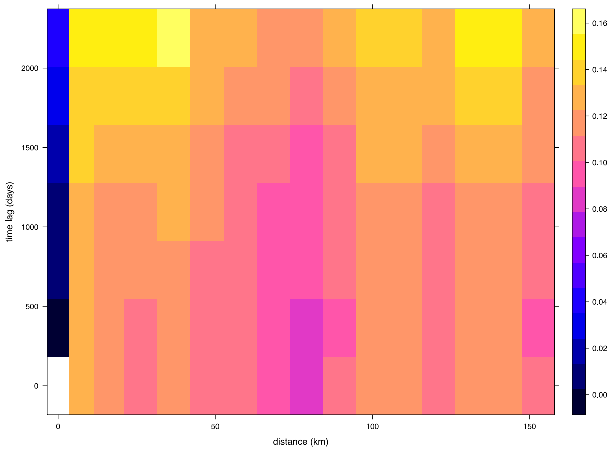


Figure 4. Empirical spatio-temporal variogram of yearly deposits from 2010 to 2016. The time lags on the y-axis are in days and the distance on the x-axis is in kilometers. The color bar indicates the value of gamma.

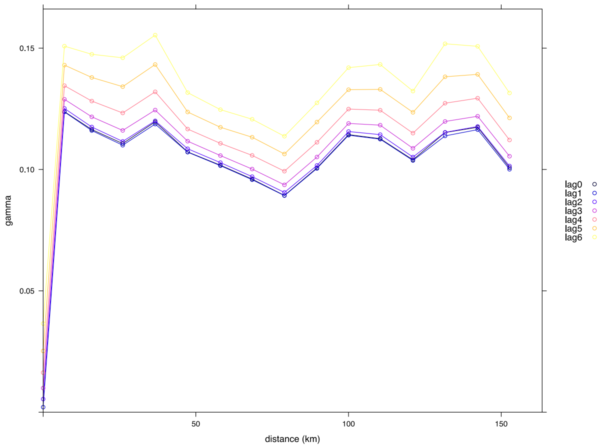
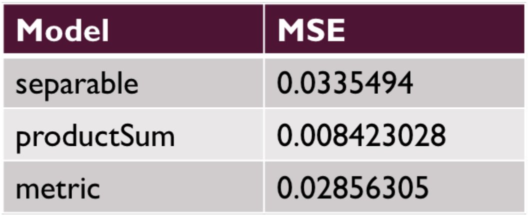


Figure 5. Empirical spatio-temporal of yearly deposits from 2010 to 2016. The gamma is on the y-axis and the distance on the x-axis is in kilometers. The color of each line represents different time lags. Here, 7 time lags are 7 years.

2.3. Model Specification

The algorithm is built to try sets of different spatio-temporal models available in gstat package. Based on comparing the mean squared error (MSE) in Table 1, the best model to fit the data is the product sum model with the lowest MSE.



Model	MSE
separable	0.0335494
productSum	0.008423028
metric	0.02856305

Table 1. Comparison of 3 models have the best MSE

2.4. Model Fitting

Let Z = {Z(s,t)} be a second order stationary spatial-temporal random field then the product sum covariance function can be written as

$$C(h,\ \mathrm{u}) = \mathrm{k}C_s(\mathrm{h})\ C_t(u) + C_s(h) + C_t(u)$$

$k > 0$

The corresponding variogram is

$$\gamma(h,\ \mathrm{u}) = (\mathrm{k}\ \cdot sill_t + 1)\gamma_s(h) + (\mathrm{k}\ \cdot sill_s + 1)\gamma_t(u)\ - k\gamma_s(h)\gamma_t(u)$$

γ_s, γ_t: spatial and temporal variograms
$h,$ u: spatial and temporal lags

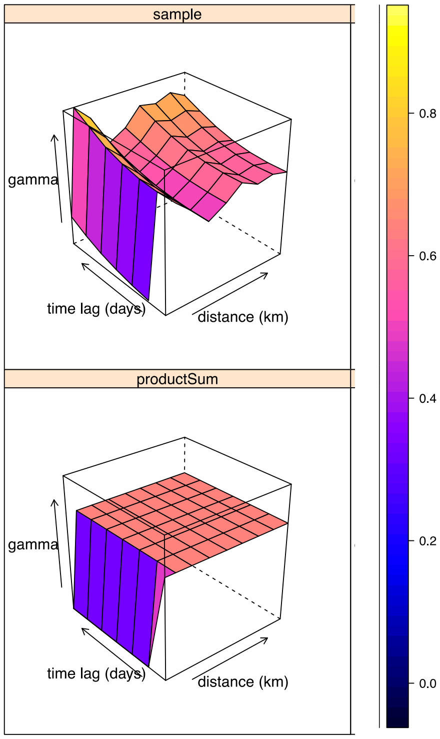


Figure 6.(Top) The empirical variogram and (bottom) the fitted variogram.

```
space component:
 model      partial sill      range
  Exp        0.6498865    50.00535

time component:
 model      partial sill      range
  Exp             0. 1          1000

k: 4.90182829588416
```

3. Results

Now, the prediction is built by using the fitted spatial-temporal covariance model and spatial-temporal universal kriging. In this case, the lattitude coordinate is treat as a covariate. The space-time prediction grid is created based on the spatial grid and temporal grid. The spatial grid considers 20 spatial locations between 85°W and 80°W, and 20 spatial locations between 38°N and 42°N. The temporal grid considers 7 equally spaced year from 2010 to 2016. The prediction of money deposits for 7 years in Figure 7 (left) are increasing each year which are consistent with the original data. The model also able to predict the amount of deposits for unobserved location which look reasonable.

The prediction standard errors map in Figure 7 (right) showing the uncertainty of the predictions. Higher certainty values are indicated in purple. Contour intervals varies from 0.65 to 0.85. The higher the standard error values, the higher the uncertainty. The area that has more bank locations result in higher certainty. Here, the prediction standard error is defined as

$$\sigma_{pred} = \sqrt{Var_{pred}}$$

σ_{pred} is the prediction standard error
Var_{pred} is the prediction variance

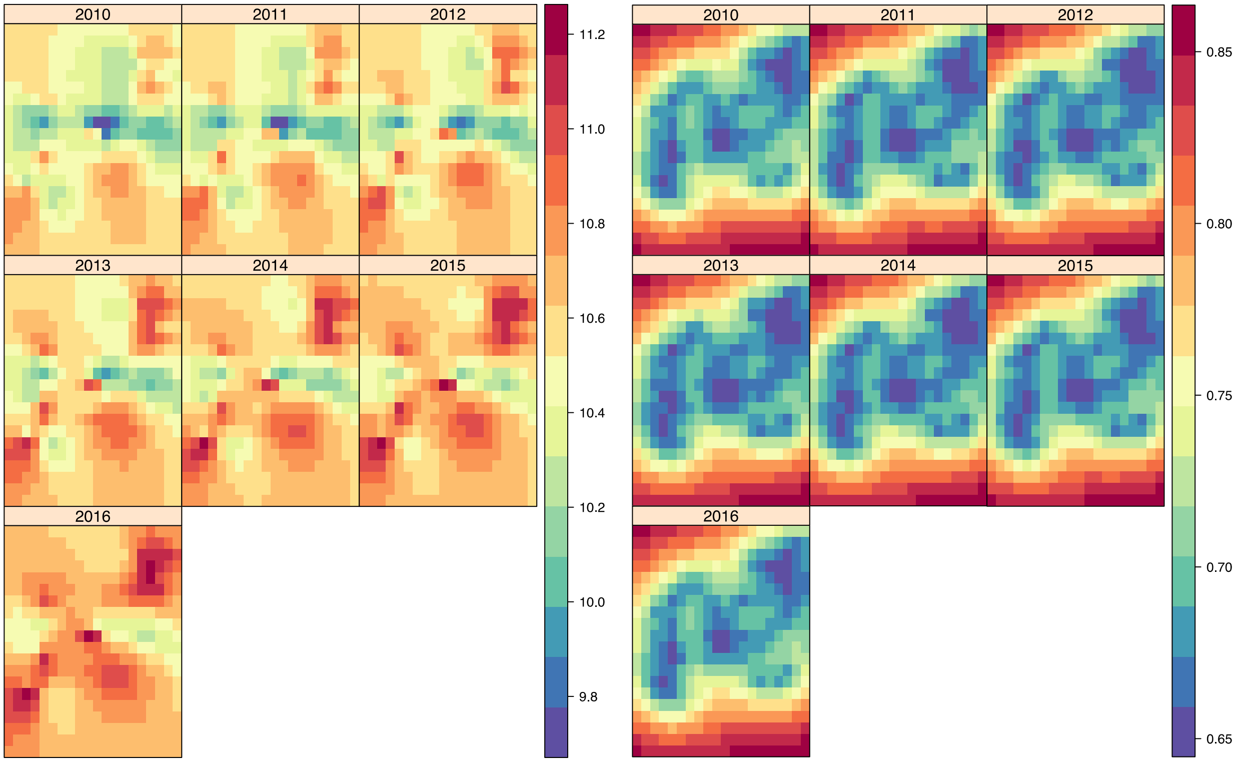


Figure 7. (Left) Spatio-temporal universal kriging predictions and (right) prediction standard errors of deposits (log $000) within a rectangle lat-lon box enclosing the domain of interest for 7 years from 2010 to 2016.

4. Discussion

From the prediction of money deposits (log $000) for 7 years, we can see the deposits are increasing from each year. Each area of Ohio state has different rate of growth of money deposits. The central area in 2010 has the deposits about 16318000 (log scale is 9.7) and in 2016 it has the deposits of 73130000 (log scale is 11.2) which means the deposits increase by a factor of 5. In comparison, the North East and South West area has the deposits which increase by factor of 2. In the central area, Columbus is Ohio state capital and has the highest population and also has the most Chase bank locations (27 braches). However, in 2010, the central area has the lowest money of deposits. It would be interesting to explore the growth rate of deposits as a geostatistics problem to find the spatial autocorrelation in the future.

References

[1] Banks around the World. (2019) Retrieved from
https://www.relbanks.com/top-us-banks/deposits.

[2] Bivand, Roger, et al. (2013). Applied Spatial Data Analysis with R. New York: Springer.

[3] FDIC. (2019). Retrieved from
https://www.fdic.gov

[4] Wikle, Christopher K., et al. (2019). Spatio-Temporal Statistics with R. Florida: CRC Press, Taylor & Francis Group.