Latent Voronoi Spatial(LVS) model
for geo-spatial reasoning

1 Motivation

Geospatial data is data about objects, events, or phenomena that have a location on the surface of the earth.[3] The widespread use of geospatial data enables lots of scientific studies and real-world applications. For example, geospatial data allows researchers to predict natural disasters[2, 6], predict forest fire[5], to study epidemiology[4], to predict house prices [1], and even crime rates [7].
However, existing geospatial prediction methods often fail to capture piecewise sharp boundaries. Therefore, the LVS(latent Voronoi Spatial) model was proposed by Ga Wu et al. to support geospatial predictions. In the LVS model, each data point defines a Voronoi cell to maintain the granularity of spatial representation. In this project, I would like to build on the LVS model to perform short-term rental price prediction with spatial features.
My research questions for this project are: (1) Can LVS model predict short-term rental prices with low RMSE error? (2) How does LVS perform compared to KNN with sparse data?

2 Model

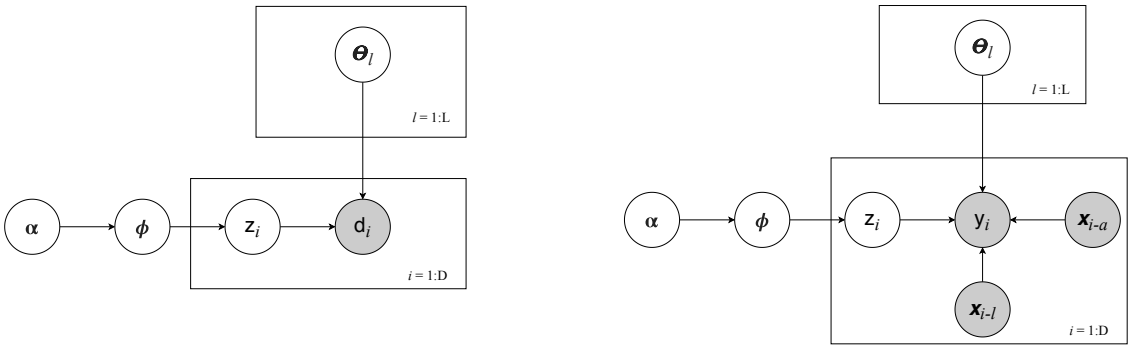

Figure 1: LVS model

Figure 1 (a) provides the graphical model of the original LVS model. The nodes represent the following:

- d_i: represents data i we observed. As shown in Figure 1 (b), we may explicitly represent $d_i = (\mathbf{x}_i; y_i)$
- z_i': represents latent variable of data point i. (the plate outside of z and d indicates size of data points, and we use $|D|$ represent number of data points in the dataset)
- ϕ: represents a multinomial distribution that z is sampled from, which define number of possible outcome of z.
- θ_l: represents linear regression parameters of latent segmentation l that defined by z.
- α: represents Dirichlet process parameter that ϕ sampled from.

To extend the existing LVS model, I introduced the graphical representation in Figure 1 (b), where individual data points are represented in features $\mathbf{x}_{i-a}, \mathbf{x}_{i-l}$ and targets y_i. In the graphical model, $\mathbf{x}_{i-l}$ represents the location of each data point, $\mathbf{x}_{i-a}$ represents attribute features. It is important to note that in geospatial prediction tasks, the location of interest $\mathbf{x}_{i-l}$ is always observed.

3 Implementation

I used pymc3 for this project. There are two implementations:

- LVS implementation with class variable z_i
- Marginalized LVS implementation without z_i

The implementation of the LVS model explicitly modelling z can be found in *model/LVS-model-withz.ipynb*.While fitting the model, I used a NUTS sampler for σ, θ, ϕ and a CategoricalGibbsMetropolis sampler for the class variable z_i. 30000 samples were drawn: two chains with 5000 tuning samples and 10000 samples for each chain.
The implementation of marginalized LVS model can be found in *model/LVS-model-marginalized.ipynb*.For sampling, I used a NUTS sampler for all variables. While fitting the model, 44000 samples were drawn: two chains with 2000 tuning samples and 20000 samples for each chain.

Model	samples	convergence	time
LVS with z_i	30000	False	50 minutes
Marginalized LVS	44000	True	4 minutes

Table 1: Caption

3.1 Data

Due to lack of qualitative dataset with piecewise boundaries found, I created synthetic data for this project. The data can be found in the *data/synthetic/* folder.
The synthetic data lie in a 10x10 grid with 3 classes, colored with green, blue and red. The true partition of the grid is shown in 2.
The partition is generated by randomly drawing 40 points in the grid with class assignment. Each Voronoi cell that contains the point is assigned with the same class as the point. Each class is associated with a set of regression parameters.

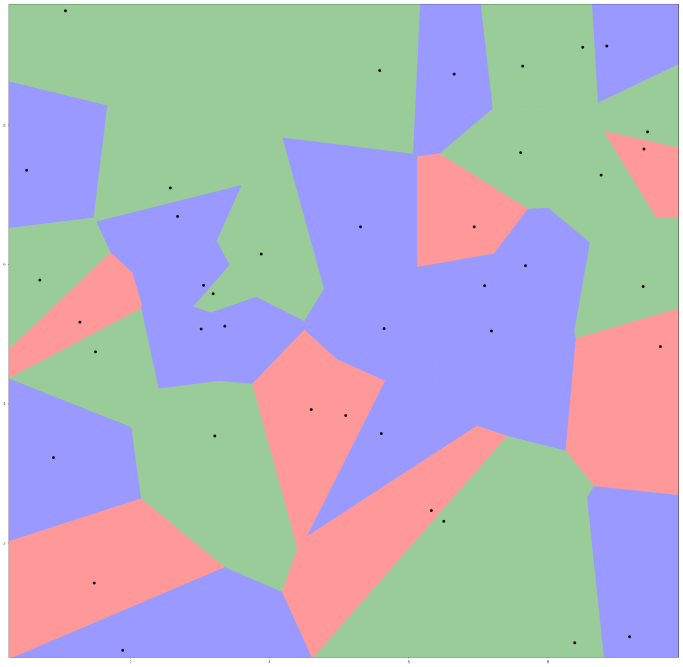

Figure 2: Ground truth partition

Then, 1000 points are randomly drawn from the 10x10 grid as training set. Each data point belongs to one of the three classes and is associated with one set of regression parameters.
The synthetic data generation can be found in */data/synthetic/synthetic_data_creation.ipynb*.

4 Experiment

4.1 LVS model with z_i

The model implementation and results for LVS model with z_i can be found in *model/LVS-model-withz.ipynb*

Data	File name	Information
Ground truth partition	synthetic-groundTruth.csv	true class, region(Multi-polygon) , true regression parameters
Synthetic training data	synthetic-3C2D.csv (1000 points)	coordinate, true class, feature 1, feature 2, true target, noisy target
Synthetic testing data	synthetic-3C2D-test.csv(200 points)	coordinate, true class, feature 1, feature 2, true target, noisy target

True area	Predicted area	Intersection area	Recall	Precision
46.50	45.30	43.00	0.92	0.95
38.22	39.03	35.42	0.93	0.91
19.31	19.68	17.06	0.88	0.87

Table 2: Area of ground truth polygon, predicted polygon and the intersection.

Reconstruction of the geospatial partition
Figure 3 shows the reconstruction of the geospatial partition after training with LVS model. The aggregated intersection area is 95.48% of the entire 10x10 grid. Table 2 contains more information of area of ground truth polygons, predicted polygons and their intersection.

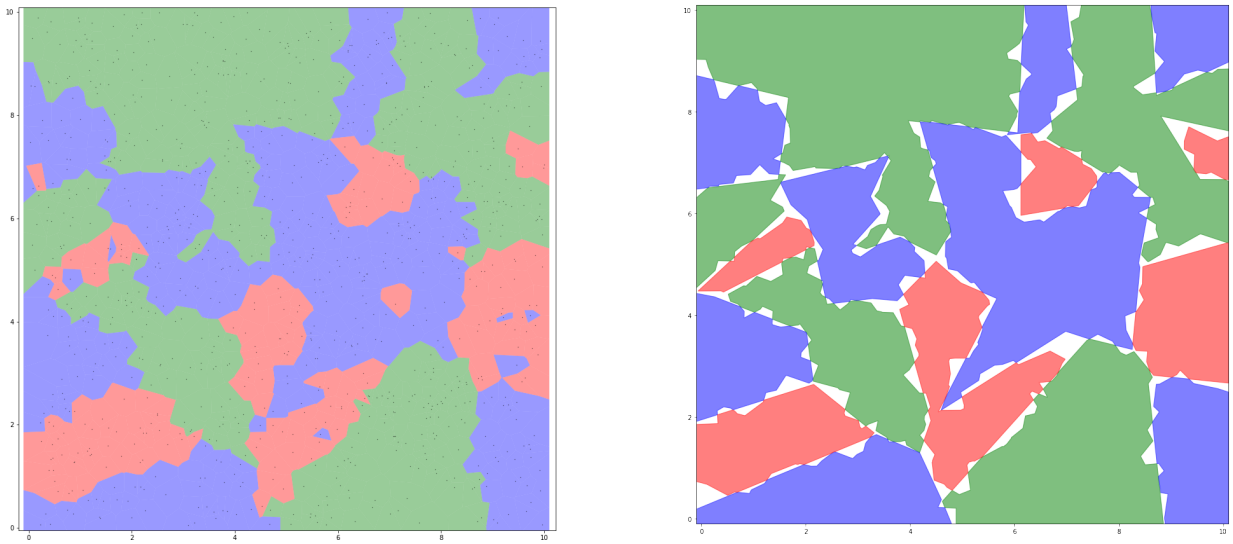

Figure 3: Left: Reconstruct of the geospatial partition with LVS mode; Right: Intersection with ground truth partition

Learned class z_i
Although the sampler did not converge, the model was still able to learn z_i very well. Figure 4 shows the confusion matrix of learned z_i.
Also, the learned z_i has very high confidence. Figure 5 shows the histogram of learned z_i.

Learned regression parameters
The regression parameters learned from the sampler was able to reproduce the true regression parameters perfectly.

Performance on Testing data
For testing, the class of testing data points was predicted according to the reconstruction of the spatial partitioning as shown in Figure 3. The predicted class z_i and target y are stored under *model/prediction*. The RMSE for test data target y prediction was 7.55. The accuracy of predicted class is 0.91.

4.2 Marginalized LVS model

The model implementation and results for LVS model with z_i can be found in *model/LVS-model-marginalized.ipynb*

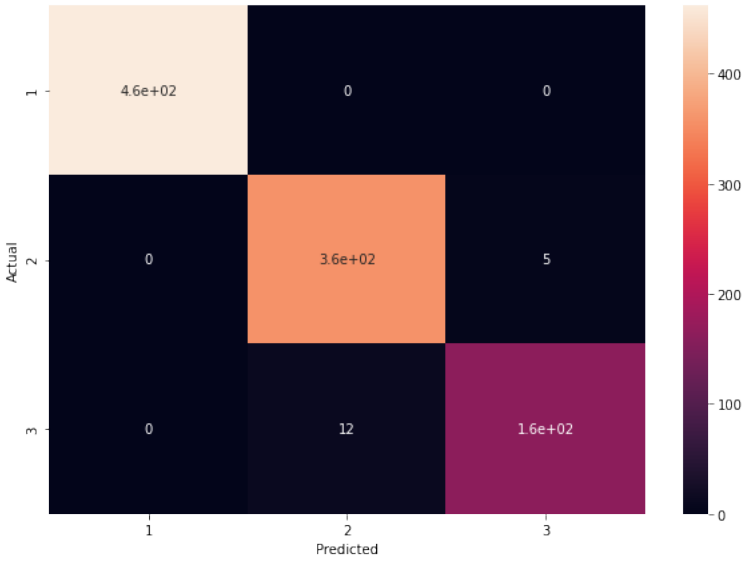

Figure 4: Confusion matrix of the learned z_i

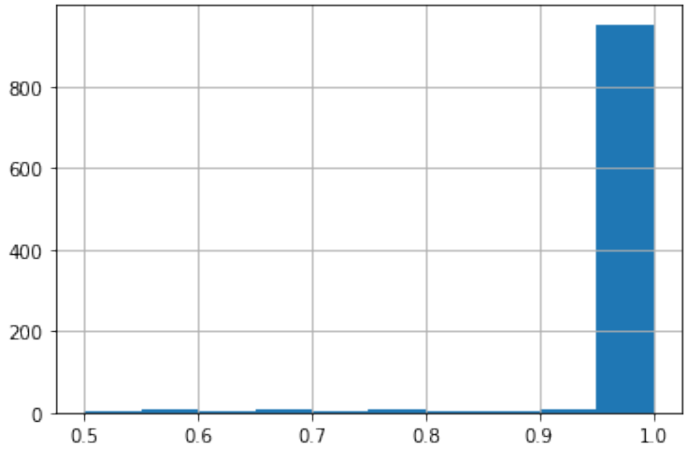

Figure 5: Histogram of confidence of learned z_i

The Marginalized LVS model was able to draw more samples a lot more faster and was also able to recover the true regression parameters. The future work of this project is to recover z_i from the marginalized model and reconstruct the geospatial partitions.

5 Discussion and Future works

From the experiments, it can be seen that LVS model has great potential to perform geo-spatial reasoning especially with sharp piece-wise boundaries.
However, the scope of this project was limited by the speed of sampling with pymc3. Future work includes:

- Work with marginalized LVS model to recover z_i.
- Implement customized faster Gibbs sampler.
- Test LVS model with real-world applications.

References

[1] Bourassa, S., Cantoni, E., and Hoesli, M. Predicting house prices with spatial dependence: A comparison of alternative methods. *Journal of Real Estate Research 32* (04 2010), 139–160.

[2] Deschamps, A., Greenlee, D., Pultz, T. J., and Saper, R. Geospatial data integration for applications in flood prediction and management in the red river basin. In *IEEE International Geoscience and Remote Sensing Symposium* (June 2002), vol. 6, pp. 3338–3340 vol.6.

[3] Layton, R., and Watters, P. *Automating Open Source Intelligence: Algorithms for OSINT*. 01 2015.

[4] Lopez, D., Gunasekaran, M., Murugan, B. S., Kaur, H., and Abbas, K. M. Spatial big data analytics of influenza epidemic in vellore, india. In *2014 IEEE International Conference on Big Data (Big Data)* (Oct 2014), pp. 19–24.

[5] Tehrany, M., Jones, S., Shabani, F., Martínez-Álvarez, F., and Tien Bui, D. A novel ensemble modelling approach for the spatial prediction of tropical forest fire susceptibility using logitboost machine learning classifier and multi-source geospatial data. *Theoretical and Applied Climatology* (09 2018).

[6] Thao, N., Hoang, N.-D., Pradhan, B., Quang Khanh, N., Truong, T., Quang Minh, N., Samui, P., and Tien Bui, D. Novel hybrid swarm optimized multilayer neural network for spatial prediction of flash flood at tropical area using sentinel-1 sar imagery and geospatial data. *Sensors* (10 2018).

[7] Zhuang, Y., Almeida, M., Morabito, M., and Ding, W. Crime hot spot forecasting: A recurrent model with spatial and temporal information. In *2017 IEEE International Conference on Big Knowledge (ICBK)* (Aug 2017), pp. 143–150.