E. Details on NC Tracts Data

E.1. *Application Details*

The full list of variables used in the clustering analysis is as follows:

Again, the analysis was performed on an Apple Macbook Pro with M1 Pro processor with 32 GB of memory. The scikit-learn clustering package[24] package was used for all experiments to perform comparison K-means clustering as well as handling Gaussian process modeling for the GPSC algorithm and computing clustering metrics. Although we are unable to release the data and auxiliary files for our real world application, the code for clustering and plotting has been submitted along with all the simulation code. For both K-means and GPSC, the full set of data shown in Table 12 including the covariates and spatial data were input into both algorithms.

Variable	Description
Spatial Data S	
LATITUDE	Latitude coordinate of population-weighted geographic center of tract
LONGITUDE	Longitude coordinate of population-weighted geographic center of tract
Covariates X	
PRFL_M	Men in professional occupation
PRFL_F	Women in professional occupation
LS_HS	Less than high school education
SINGLE	Single with dependent
HSHLDR_F	Female head of household
NHBLK	Non-Hispanic Black
PA	Public assistance
POV	Poverty
NO_VHCL	No vehicle
RENT	Rental housing
CROWD	Crowded housing
UNMPLYD	Unemployment
PHONE	No phone
ACET	Acetaldehyde
BENZENE	Benzene
BUTA	1,3-Butadiene
CARBON	Carbon Tetrachloride
DIESEL	Diesel PM2.5
ETHYL	Ethylbenzene
FORM	Formaldehyde
HEXANE	Hexane
LEAD	Lead compounds
MANG	Manganese compounds
MERC	Mercury compounds
METH	Methanol
METHYL	Methyl Chloride
NICK	Nickel
TOLUENE	Toluene
XYLENE	Xylenes
Response Y	
MLCJOINT	Overall class membership into 8 possible groups

Table 12: Full set of variables used for NC tracts data application.

E.2. *Additional Real World Application Comparisons*

In this section we present additional results of the best performing competitors from the simulation studies (using default parameters) on the CBCS real world example of the main paper, which already contained the comparison to K-means clustering. We also include one algorithm of each type from the set of competitors, again for diversity of results. It can be seen that the different types of clustering models have distinct differences to the results of GPSC as discussed below.

E.2.1. *Gaussian Mixture Model*

Here we report the clustering results of the Gaussian Mixture Model. It can be seen that the results are visibly similar to the results of K-means clustering, where again the algorithm appears to mostly center the cluster diversity around the major urban centers of the state, with fewer cluster diversity across the extremities and regions between the urban centers.

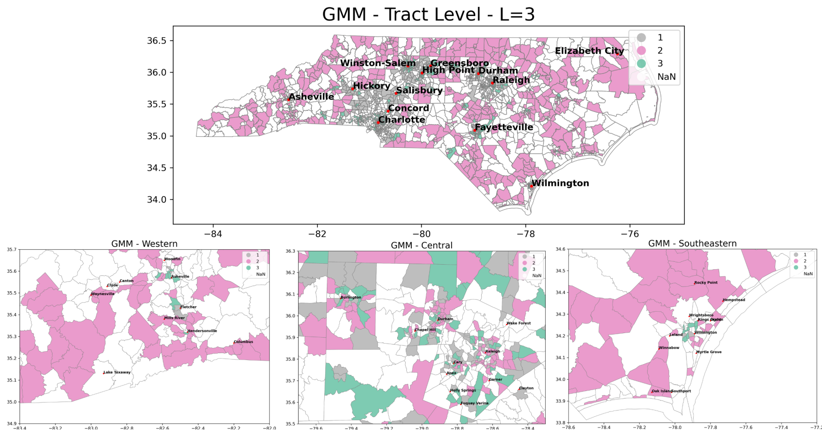

Fig. 29: GMM results for the real world application presented in main paper.

E.2.2. *Spectral Clustering*

Spectral clustering, similar to the spatial hierarchical clustering results presented below, seems to pick up more global trends with lower nuance specifically around the dense city regions of the state.

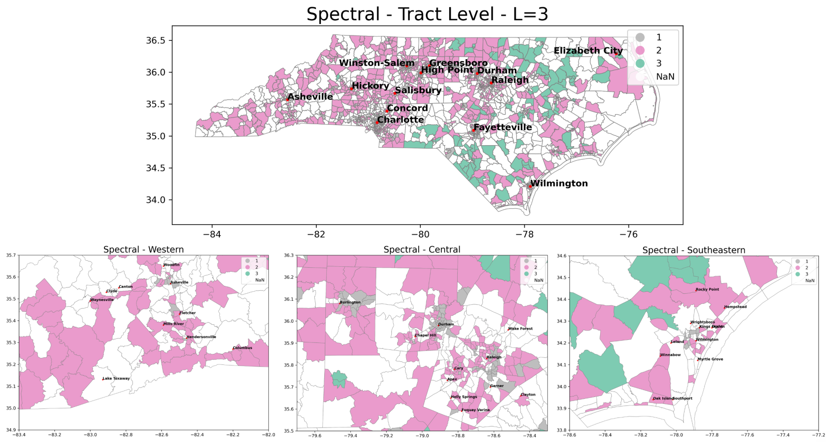

Fig. 30: Spectral clustering results for the real world application presented in main paper.

E.2.3. *Spatial Hierarchical Clustering*

Spatial hierarchical clustering is presented here with 5 neighbors (result did not vary significantly over different specifications of the neighbor count). It can be seen that although the algorithm may be picking up on more global trends across the state, there is decreased nuance around the specific city centers of the state.

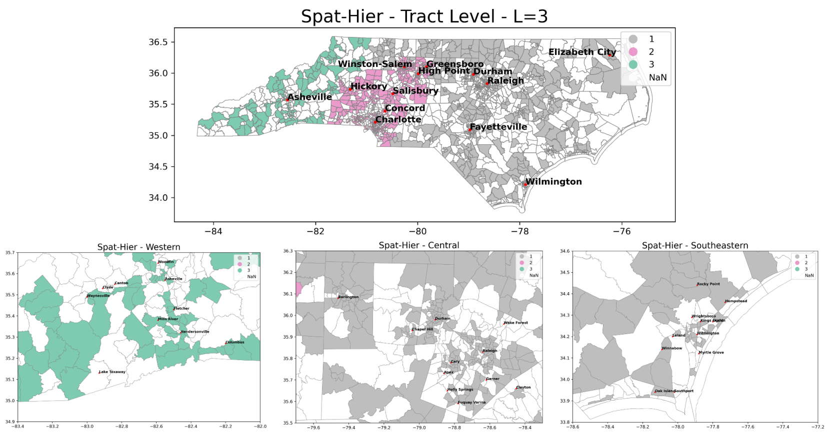

Fig. 31: Spatial hierarchical clustering results for the real world application presented in main paper.

E.2.4. *DBSCAN*

Here DBSCAN was chosen over GDBSCAN due to having fewer hyperparameters required to tune (default used), while having similar performance in the simulation studies. It can be seen here that the main challenge of DBSCAN (as well as GDBSCAN) is the inability to mandate the number of clusters, especially in this application where we specifically seek a small number of clusters for interpretability.

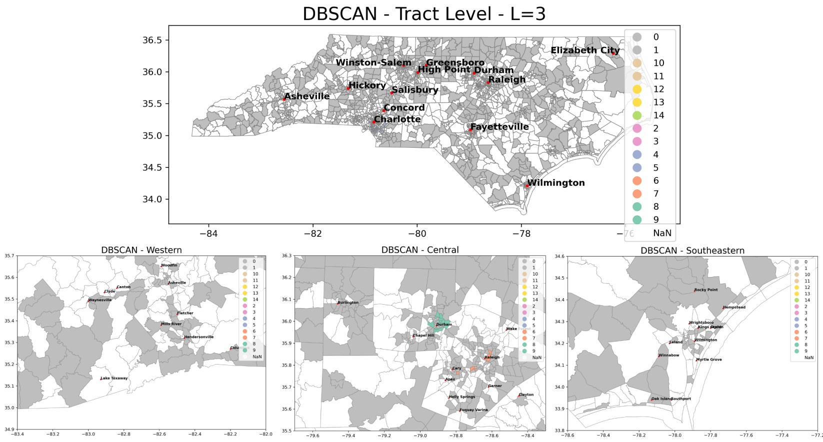

Fig. 32: DBSCAN clustering results for the real world application in the main paper.