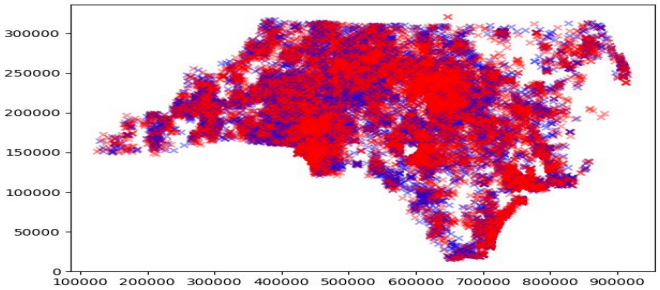
Marriage and Statistics through Space and Time

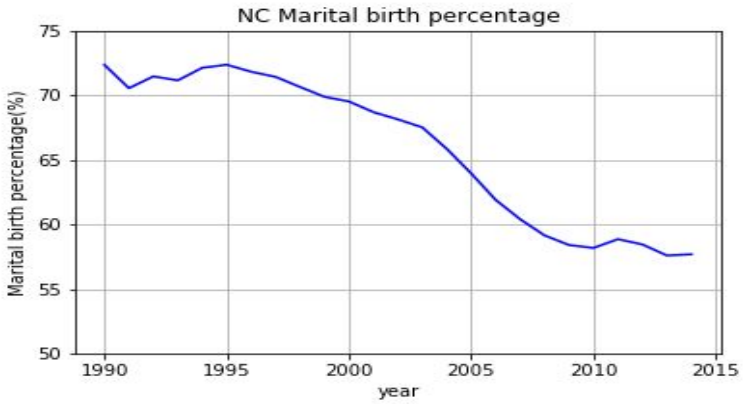
Team6: Feixiao Chen, Khuong Do, Jason Law
Project Manager: Lizzy Huang
Faculty Leads: Christina Gibson-Davis, Paul Bendich
Outside Consultants: Abe Smith, Tim Stallman

Locations of white marital (red) and non-marital (blue) births in NC, 2009



Objective

- We want to find if the distribution of births to married vs unmarried mothers is changing over time, and if the distribution of births by race is changing over time. We also want to see how these different distributions compare to each other.
- We are also interested in how this analysis of marriage and race relates to a similar analysis of poverty, education, income, etc.



Data

- Subset 2 million+ birth records due to mother marital status, races from North Carolina Birth Record Database.
- Socioeconomic indicators prepared from US Decennial Census, American Community Survey 5-Year Estimates, all redistricting to 2010 census block group geographical level for North Carolina.
  - ❏ Income: Median/State Median household income.
  - ❏ Poverty: Population below poverty line/Total population.
  - ❏ Education: % individuals with a degree above or equal to Bachelor's, and individuals without High School diploma.

References

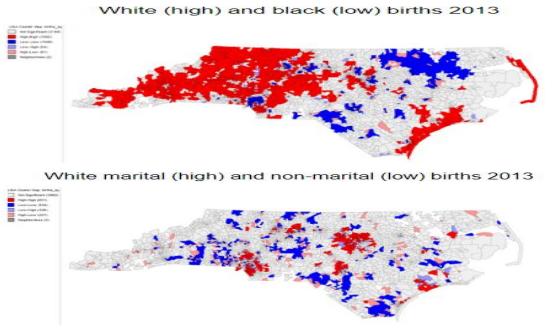
❏ Census/ACS data retrieved from https://www.nhgis.org/.
❏ Anselin, L.(1995), Local Indicators of Spatial Association - LISA.
❏ Smith, A., Bendich, P., Harer, J. and Hinema, J. (2017), Supervised Learning of Labeled Pointcloud Differences via Cover-Tree Entropy Reduction.

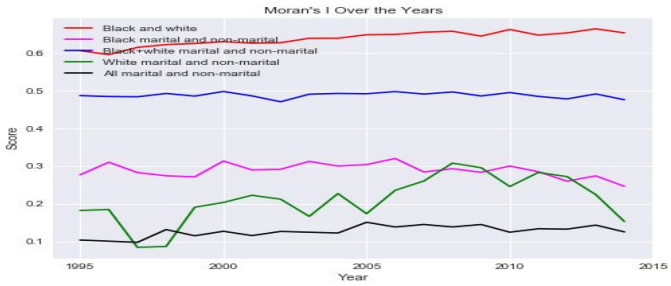
Methods

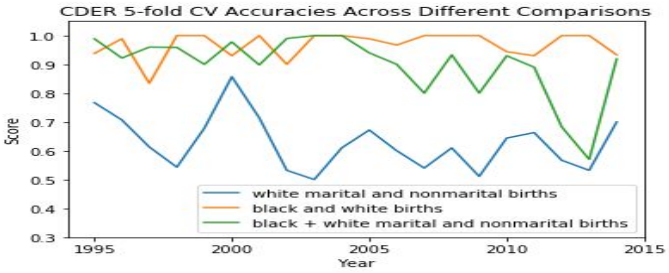
Spatial Statistics Method: Local and Global Moran's I

- **Local Moran's I** measures the spatial clustering of a numerical attribute (e.g proportion of white births in a given block group).
  - ❑ A continuous variable, but could be classified into four types/colors: Red = high-high (a high-proportion block group is surrounded by other high-proportion block groups). Blue = low-low. Light blue = low-high. Pink = high-low.
  - ❑ Lots of red and blue signify lots of clustering of like values (see references).
  - ❑ High refers to "high proportion of married (compared to unmarried) births" or "high proportion of white (compared to black) births.
  - ❑ Results: Local maps show that black/white births are more clustered with respect t like values than white marital/non-marital births.
- **Global Moran's I** is a global index proportional to the sum of local scores. Range = [-1, 1].
  - ❑ Positive score means that overall, similar values are clustered.
  - ❑ Results: Clustering is consistently positive and slightly increases over time for both black/white and white marital/non-marital comparisons, but the former is more strongly "clustered", confirming the local maps.

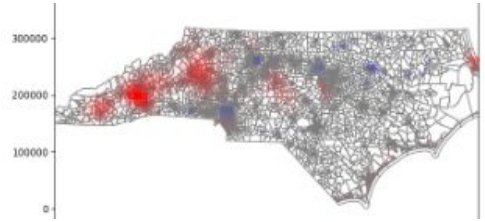
GeoDa maps of local scores for NC block groups, for white marital/non-marital (top) and white/black (bottom) comparisons in 2009







Heatmap created from the training data for 2009 black (blue label) white (red label) births



Machine Learning Method: Point Cloud-based Machine Learning (CDER)

Point cloud classification: Cover-tree Differencing via Entropy Reduction (CDER):

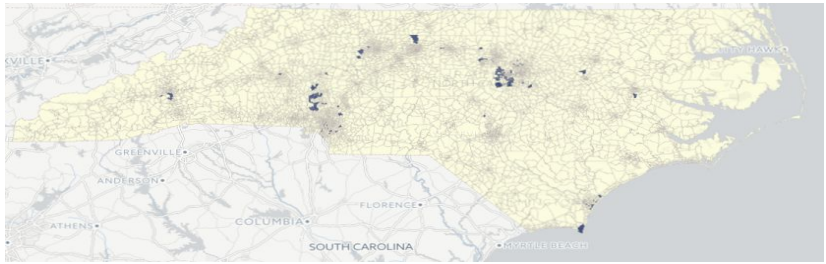
- Task: classify a point cloud of birth coordinates as "marital" or "nonmarital"
- CDER is a new machine learning method that labels point clouds using information theory and computational geometry (see references).
- Applied to annual birth data: Split each year into 12 months * 4 weeks/month = 48 weeks: each week gives 2 point clouds of different labels (e.g. marital and non marital, black and white). Train 80% of the week, creating Gaussians to predict the test weeks. If CDER does a good job predicting, then the distributions of the two labels are very different.
- Results: CDER accuracy is consistently higher for the black-white comparison than the white marital-nonmarital comparison, suggesting a stronger spatial distinction between the labels that correspond to the former.

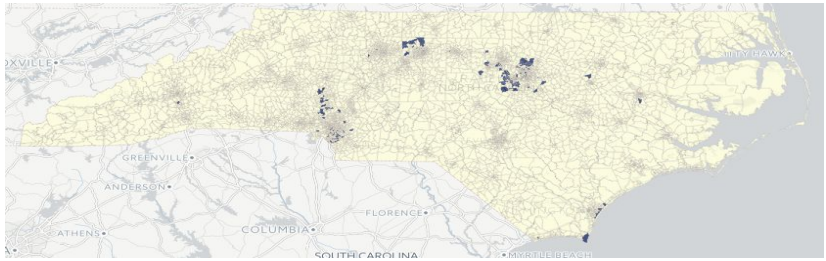
Socioeconomic Analysis

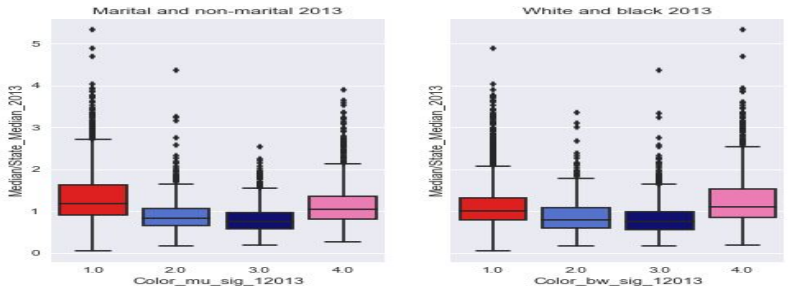
R Shiny app - dataplus2017.shinyapps.io/Dataviz

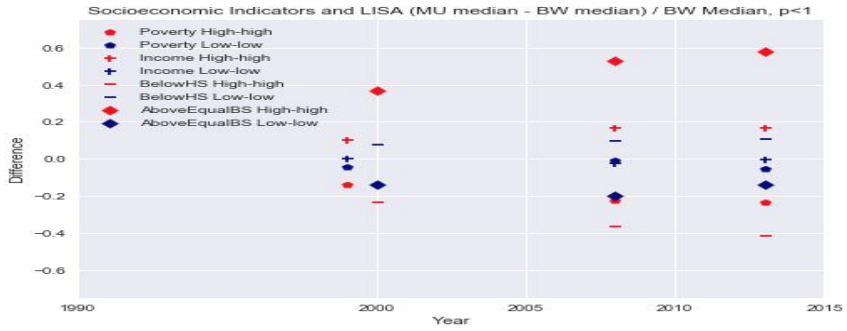
Created an R Shiny app that allows geographical visualization of the variables used and filtering by user-specified parameters.

- For example, the top plot shows the block groups which are high-high white block groups that have a median income that is at least double the state median income in 2013. These block groups can be interpreted as wealthy white clusters. There are 125 such block groups.
- For comparison, the bottom plot is the same plot, but for high-high marital block groups, meaning wealthy married clusters. There are 181 such block groups.









Local Moran's I versus Socioeconomic Variables

- For ~6000 block groups, plot the "median/state median household income" with LISA colors. Then compare the median income of very married clusters (left-red box) to that of very white clusters (right-red box). **In 2013, being married shows a higher median income level than just simply being in a white group.**

- The scatter plot on the left shows the *percentage differences (0.2 = 20%)* between the same-colored medians for corresponding "marriage" and "race" comparisons, for income, poverty and education, and for years 2000, 2008, and 2013. The scatter trend is consistent with the hypothesis: **being in a high-high marital cluster consistently gives a more desirable socioeconomic level than being in a high-high white cluster. Also, the advantage seems most pronounced for education.**