Paper PO-14

Spatial Analysis of Gastric Cancer in Costa Rica using SAS

So Young Park, North Carolina State University, Raleigh, NC
Marcela Alfaro-Cordoba, North Carolina State University, Raleigh, NC

ABSTRACT

Stomach cancer, or gastric cancer, refers to cancer arising from any part of the stomach. It causes about 800,000 deaths worldwide per year. Gastric cancer (GC) is the leading cause of cancer-related mortality in Costa Rican males. After breast cancer, it is the second highest cause of cancer mortality in women in Costa Rica. Most predictor variables have been based on epidemiological and social factors, yet spatial factors have not been commonly accounted in the analysis.

In epidemiology, the prevalence of a health-related state (in this case, GC) in a statistical population is defined as the total number of cases in the population, divided by the number of individuals in the population. It is used as an estimate of how common a disease is within a population over a certain period of time. It helps health professionals understand the probability of certain diagnoses and is routinely used by epidemiologists, health care providers, government agencies and insurers. The objective of this analysis is to identify if there exists a spatial variation of GC in Costa Rica after accounting social factors, and to propose a possible methodology that can be imposed to improve future public health efforts for decreasing such prevalence. Using SAS, we construct a geographical representation of variables by county, and analyze residuals from a regression that links the social factors to GC to test the existence of spatial correlations.

INTRODUCTION

In the past, most of statistical data analysis was done using a multivariate regression model without taking account of geographical variables. However, as georeferenced data has become more accessible, including a spatial factor and explaining phenomena of interest using a spatial factor has allowed us to explain a response variable from a different perspective. The objective of this paper is to illustrate how to read data in a proper format for spatial analysis, to construct maps and to test for an existence of spatial variation of the response variable, i.e. gastric cancer (GC).

READING THE DATA

PROC MAPIMPORT is to convert a shape file to a dataset FP_map, which will be used to generate a geographical representation of the response variable RP.

```
Proc mapimport out=FP_map datafile="C:\Users\costarica.shp";
Rename DCODE= PT;
run;
```

The following code is used to import data that includes RES, which is used to calculate the response variable RP, and the explanatory variables IEV, IBM and IC. The descriptions of the variables are explained in the next section.

```
Data dat;
input n NAME1 $ NAME2 $ PT codccp born res pop IEV IBM IC;
datalines;
75 SANJOSE SanJose 48007015 101 86 85 333876 0.822621 0.688346 0.895533
77 SANJOSE Escazu 48007007 102 2 6 58121 0.848121 1 0.916214
87 SANJOSE Desamparados 48007005 103 7 48 239798 0.794035 0.579381 0.887462
86 SANJOSE Puriscal 48007014 104 7 9 30867 0.886778 0.38499 0.939588
99 SANJOSE Tarrazu 48007017 105 4 5 15789 0.85501 0.308365 0.834189
90 SANJOSE Aserri 48007003 106 7 11 53466 0.787383 0.496851 0.84244
82 SANJOSE Mora 48007011 107 1 2 25038 0.913241 0.551856 0.892494
71 SANJOSE Goicoechea 48007008 108 2 27 127140 0.858218 0.622057 0.894688
[omitted]
95 LIMON Talamanca 48005006 704 0 4 30573 0.64621 0.408296 0.635427
41 LIMON Matina 48005003 705 0 1 41271 0.744152 0.379599 0.779796
33 LIMON Guacimo 48005001 706 0 3 42426 0.870487 0.392589 0.796543;
```

We create new data called FINAL by merging two datasets, Fp_map and Dat, by the identifier PT. Because some of counties in Costa Rica have more than two separate areas, we assigned a different identifier value that distinguishes those areas in order to create an appropriate map. We also eliminate islands and lakes, -Isla and Lago Arenal-.

The response variable RP is prevalence of the gastric cancer GC, which is defined as res*10000/pop, where res is reported new cases of GC in each county during 2005 and pop is the population in county during 2005.

```
PROC SORT data = dat;
By PT;
Run;

PROC SORT data = Fp_map;
By PT;
run;

DATA cancer;
Merge Fp_map dat;
By PT;
Run ;

Data temp;
Set cancer;
if lacar2_ EQ 4424 then pt=1 ;
if lacar2_ID EQ 4151 then pt=2 ;
if lacar2_ID EQ 4341 then pt=3 ;
if lacar2_ID EQ 4234 then pt=4 ;
if lacar2_ID EQ 4343 then pt=5 ;
if lacar2_ID EQ 4308 then pt=6 ;
if name2 in ('Isla', 'lag.arenal') then delete;
Run;

Data final;
Set temp;
rp = res*10000/pop;
Run;
```

DESCRIPTION OF THE DATA

The gastric cancer morbidity data by county of residence from 1981 to 2005 is available on the Costa Rican National Cancer Records, provided by the Central American Center of Population (www.ccp.ucr.ac.cr). The Human Development Index data by county from 1992 to 2006 is available on the Observatory of Development from the University of Costa Rica (www.tdc.odd.ucr.ac.cr). In addition, population projections from 1981 to 2005 are available at the Central American Population Center (www.ccp.ucr.ac.cr).

The following list is the descriptions of the response variable and the explanatory variables used in the analysis.

- Response variable RP: Prevalence of gastric cancer RP = $\frac{Res}{Pop} * 10000$, where Res is the reported new cases of GC in a county during 2005 and Pop is population in the corresponding county
- IEV is Health Index from HDI for the corresponding county in 2005
- IBM is Education Index from HDI for the corresponding county in 2005
- IC is Income Index from HDI for the corresponding county in 2005

GEOGRAPHICAL REPRESENTATION OF THE RESPONSE VARIABLE RP

```
goptions
reset=all;
Proc gmap data= final_adj_cntr map=final_adj_cntr;
Id fprt;
choro rp;
run;
quit;
```

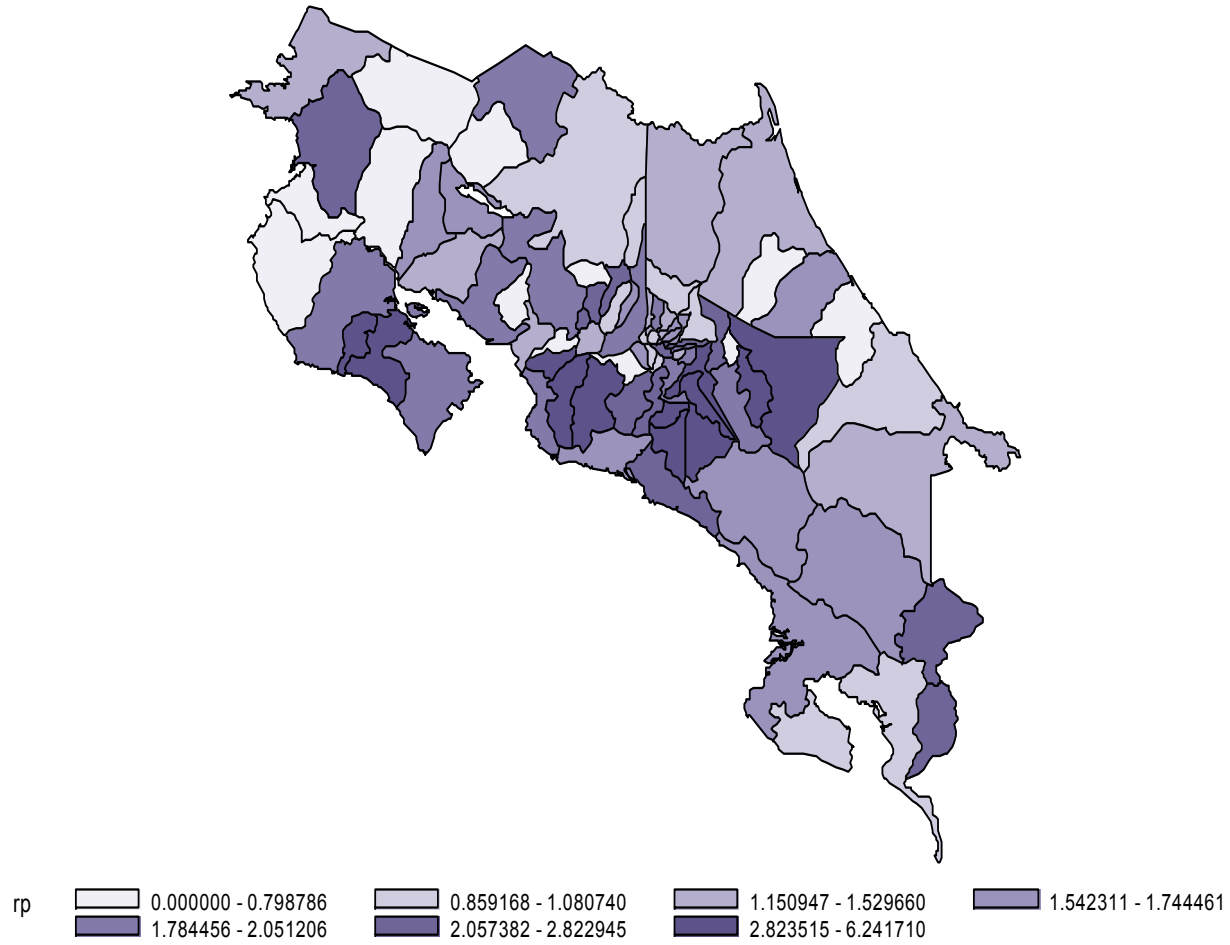

Figure 1. Prevalence of Gastric Cancer by county in Costa Rica in 2005

The presence of some clusters or accumulation of counties with high prevalence in the same geographical area can be noted in the map. Therefore it is important to test if the spatial pattern is statistically significant, or in other words, if there exists a spatial covariance structure in the data, after we explain the response variable with social factors as covariates.

STATISTICAL MODEL

An ordinary least square multiple regression model is fitted to gastric cancer prevalence (RP) using the three explanatory variables (IEV, IBM and IC). The following output gives ANOVA table and the estimates of the parameters in OLS regression model.

```
ods
graphics on;
title  "Simple Linear Regression with Diagnostic Plots";
```

```
Proc reg data=data_adjac_cntr;
model rp= IEV IBM IC / stb clb;
OUTPUT OUT=OUTREG1 P=PREDICT R=RESID RSTUDENT=RSTUDENT COOKD=COOKD;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: rp

Number of Observations Read	86
Number of Observations Used	86

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	10.04366	3.34789	3.75	0.0141
Error	82	73.29477	0.89384		
Corrected Total	85	83.33842			

Root MSE	0.94543	R-Square	0.1205
Dependent Mean	1.75933	Adj R-Sq	0.0883
Coeff Var	53.73820		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-1.29372	2.14913	-0.60	0.5489	0
IEV	1	-2.13549	2.47647	-0.86	0.3910	-0.10100
IBM	1	-2.07705	0.73352	-2.83	0.0058	-0.34715
IC	1	6.83516	2.37337	2.88	0.0051	0.37116

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	-5.56902	2.98159
IEV	1	-7.06198	2.79099
IBM	1	-3.53625	-0.61785
IC	1	2.11377	11.55655

Output 1. Output from a REG Procedure

Residuals are obtained from the OLS multiple regression model. The fit diagnostics of the response variable RP are shown in display 2. Residuals were normally distributed with p-value < 0.1. However, we suspect the presence of heteroschedasticity from the residual plot. It is violation of homogeneity of variances, one of the assumption for an OLS multiple regression. We continue the analysis to see whether this violation is due to the presence of spatial variation in gastric cancer prevalence in Costa Rica during 2005.

```
Title  "Simple Linear Regression with Diagnostic Plots";
Proc reg data=dat;
Model rp= IEV IBM IC / stb clb;
OUTPUT OUT=OUTREG1 P=PREDICT R=RESID RSTUDENT=RSTUDENT COOKD=COOKD;
Run;
Quit;
```

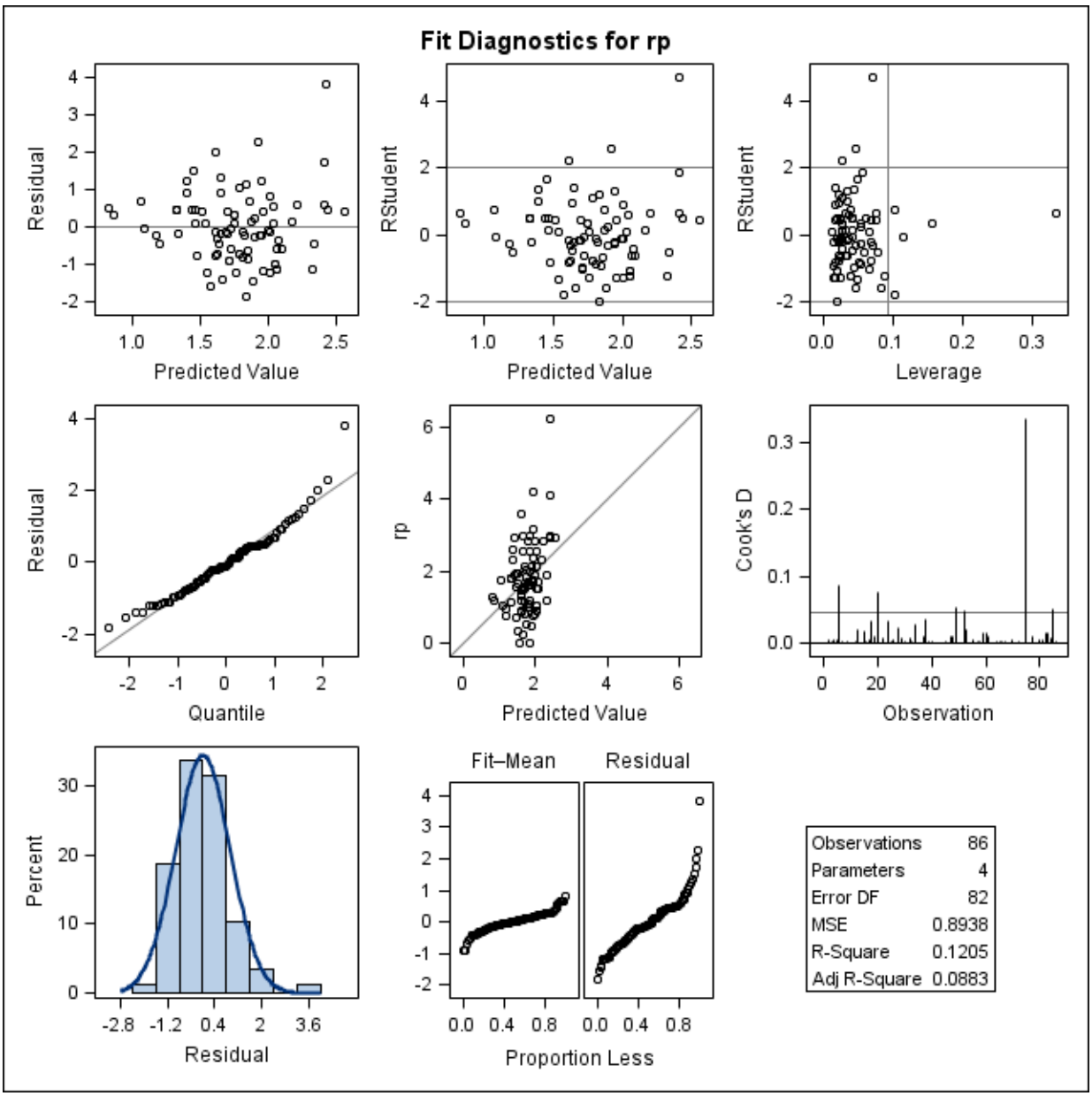

Figure 2. Diagnostics of Gastric Cancer Prevalence for the OLS model

We generate a map with residuals obtained from OLS model to see whether there is a spatial variation in residuals. In order to obtain a map, we generate a new dataset Final_Res, which includes both coordinates and residuals.

```
PROC SORT data =final;
By n;
Run;
Quit;

PROC SORT data=outreg1;
By n;
Run;
Quit;
```

```
Data final_res;
Merge final outreg1;
By n;
Run;

goptions
reset=all;
Proc gmap data= final_res map=final_res;
Id n;
Choro resid;
Run;
Quit;
```

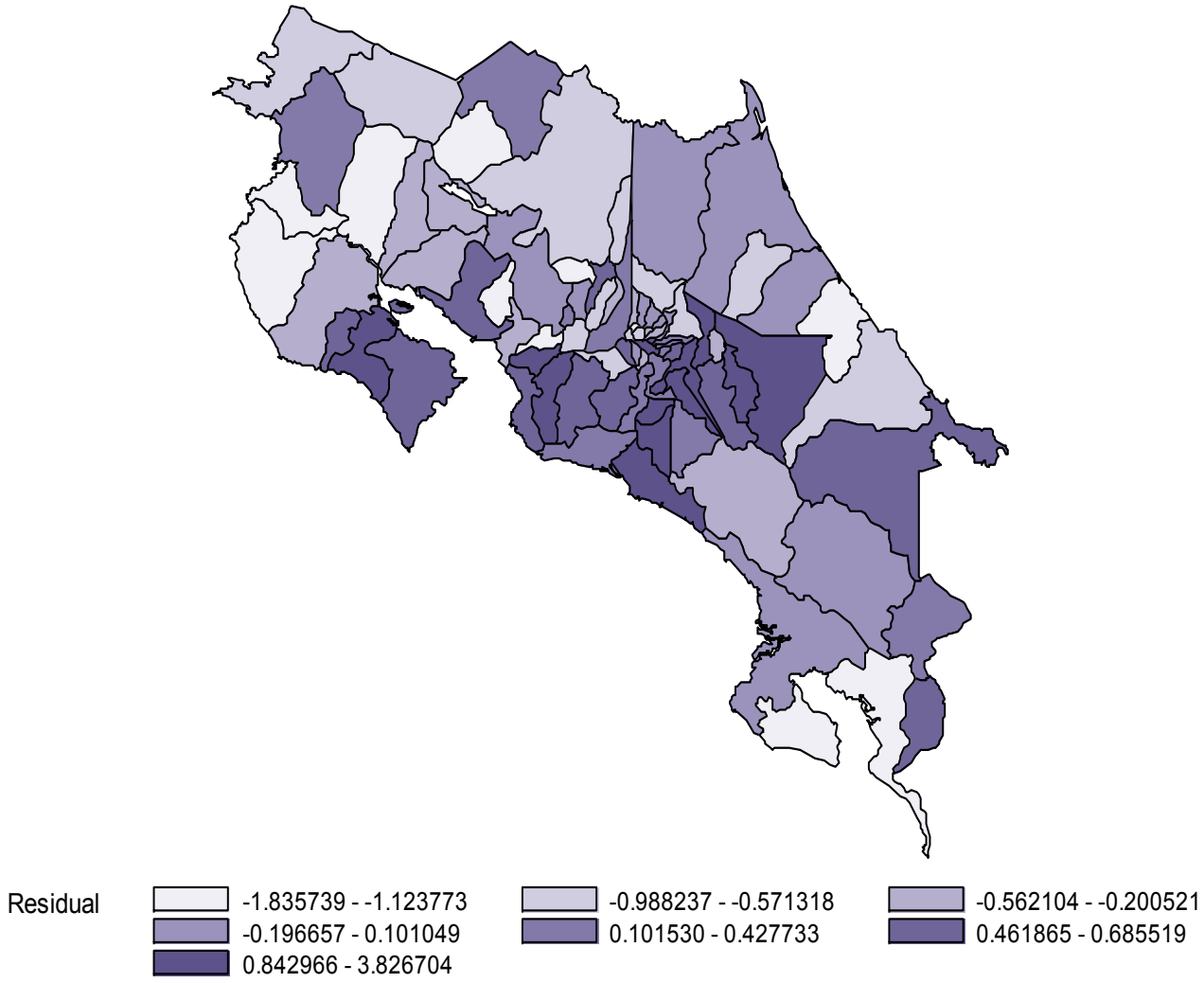

Figure 3. Residuals from OLS by County in Costa Rica during 2005

Again, the presence of some clusters or accumulation of counties with high prevalence in the same geographical area can be noted in the map. The next step then, is to test if the spatial pattern is statistically significant on the residuals, using Moran's Test.

MORAN'S I TEST

As we suspect the presence of spatial variation in residuals, we conduct Moran's I test to see the significance of the spatial variation in explaining the response variable RP. As shown in the output below, there is significant evidence that there exists the spatial variation in residuals with p-value < 0.001.

The assumption of normality is supported by the diagnostics plot from the previous section, in which we can see that most of the observed residuals where on the line of normal quantiles.

```
Proc variogram data=final_res;
Compute  novar autoc (weights=distance);
Coordinates xc=x yc=y;
Var rp;
Run;
```

The VARIOGRAM Procedure
Dependent Variable: rp

Number of Observations Read	6214
Number of Observations Used	6214

Pairs Information

Number of Lags	11
Lag Distance	0.47
Maximum Data Distance in X	3.39
Maximum Data Distance in Y	3.18
Maximum Data Distance	4.65

Pairwise Distance Intervals

Lag Class	Bounds		Number of Pairs	Percentage of Pairs
0	0.00	0.23	844276	4.37%
1	0.23	0.70	4.09E6	21.18%
2	0.70	1.16	4.75E6	24.62%
3	1.16	1.63	4.11E6	21.31%
4	1.63	2.09	2.68E6	13.86%
5	2.09	2.56	1.54E6	7.98%
6	2.56	3.03	852781	4.42%
7	3.03	3.49	364632	1.89%
8	3.49	3.96	67224	0.35%
9	3.96	4.42	3445	0.02%
10	4.42	4.89	0	0.00%

Autocorrelation Statistics

Assumption	Coefficient	Observed	Expected	Std Dev	Z	Pr > |Z|
Normality	Moran's I	0.0257	-0.000161	0.0000644	401.36	<.0001
Normality	Geary's c	1.0146	1.000000	0.0021151	6.92	<.0001

Output 2. Output from a Variogram Procedure

CONCLUSION

- Regression model is useful to describe the relationship between GC and social factors, but it fails to fulfill the assumptions, because the variance is not constant in the residuals.
- There is significant evidence that there exists a spatial variation pattern in residuals with p-value < 0.001. This proves that the heteroschedasticity in the residuals is due to an existing spatial pattern.
- The variable of Health Index (H) for county X in 2005 is not statistically significant in the model.
- The higher Education Index a county has, the lower prevalence of Gastric Cancer that county has on average. Also, the higher Income Index the county has the higher prevalence of gastric cancer during 2005.

FUTURE RESEARCH

For future research, it would be interesting to study the spatial pattern for the response variable, using models like CAR or SAR, with counties as the geographical unit. These models are not included in the SAS options, and need to be programmed from scratch.

Another option to study the response variable is relate it with hospitals as spatial points, instead of counties. Thus, the spatial model could be done as a spatial point pattern, instead of areas, if data is available.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Marcela Alfaro-Cordoba
North Carolina State University
malfaro@ncsu.edu

So Young Park
North Carolina State University
spark13@ncsu.edu