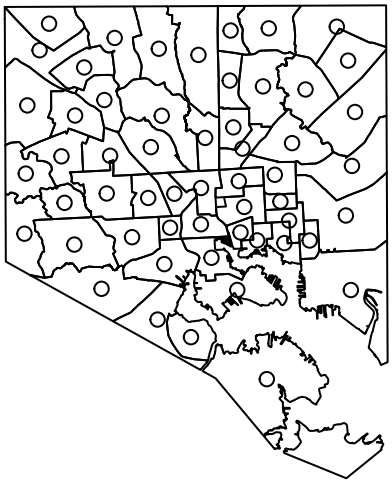
Baltimore Data Info

Introduction

Exploratory Analysis of Spatial Data Analysis & Spatial autocorrelation in selection of Models

Baltimore Map

```
## OGR data source with driver: ESRI Shapefile
## Source: "/Users/ckineza/Desktop/Census_Info", layer: "VS14_Census"
## with 55 features
## It has 28 fields
```



"Everything is usually related to all else but those which are near to each other are more related when compared to those that are further away" Tobler's First Law of Geography 1 Spatial autocorrelation is the formal property that measures the degree to which near and distant things are related

*Why is spatial autocorrelation important?

One of the main reasons why spatial auto-correlation is important is due to the fact that statistics relies on observations being independent from one another. If autocorrelation exists in a map, then this violates the fact that observations are independent from one another. Another possible reason why it is important to worry about autocorrelation of residuals is because it biases the standard errors (t values, F tests etc) of OLS estimates which meaning the model could show that variables are (in)significant when they are not Positive spatial auto-correlation occurs when Moran's I is close to +1. This means values are clustered together. Negative spatial autocorrelation occurs when Moran's I is near -1. A value of 0 for Moran's I typically indicates no autocorrelation.

Possible solutions in fixing autocorrelation in residuals:

-Adding/deleting variables -Increasing the temporal period -Adjusting the errors by first differencing and multiplying by the autocorrelation coefficient

If none of these "simple" solutions work, the methods become increasingly complex and in at least some cases, the "cure" can be worse than the "disease" it is attempting to fix

Note that addressing autocorrelation is not the full story — this is still a regression problem, and all of the usual checks (scatter plots, residual plots, diagnostics, etc.) are still essential.

Model 1: Life Expectancy ~ income, poverty and Tobacco Reading shapefile (extension is .shp, using Census.shp for city of Baltimore)

```
## OGR data source with driver: ESRI Shapefile
## Source: "/Users/ckineza/Desktop/Census_Info", layer: "VS14_Census"
## with 55 features
## It has 28 fields
```

Figure 1 : Moran Test Scatterplot: The four quadrants in the graph provide a classification of four types of spatial autocorrelation: high-high (upper right), low-low (lower left), for positive spatial autocorrelation; high-low (lower right) and low-high (upper left), for negative spatial autocorrelation. The slope of the regression line is Moran's I. 2

```
## Error in `rownames<-`(`*tmp*`, value = c("0", "1", "2", "3", "4", "5", : length of 'dimnames' [1] no
```

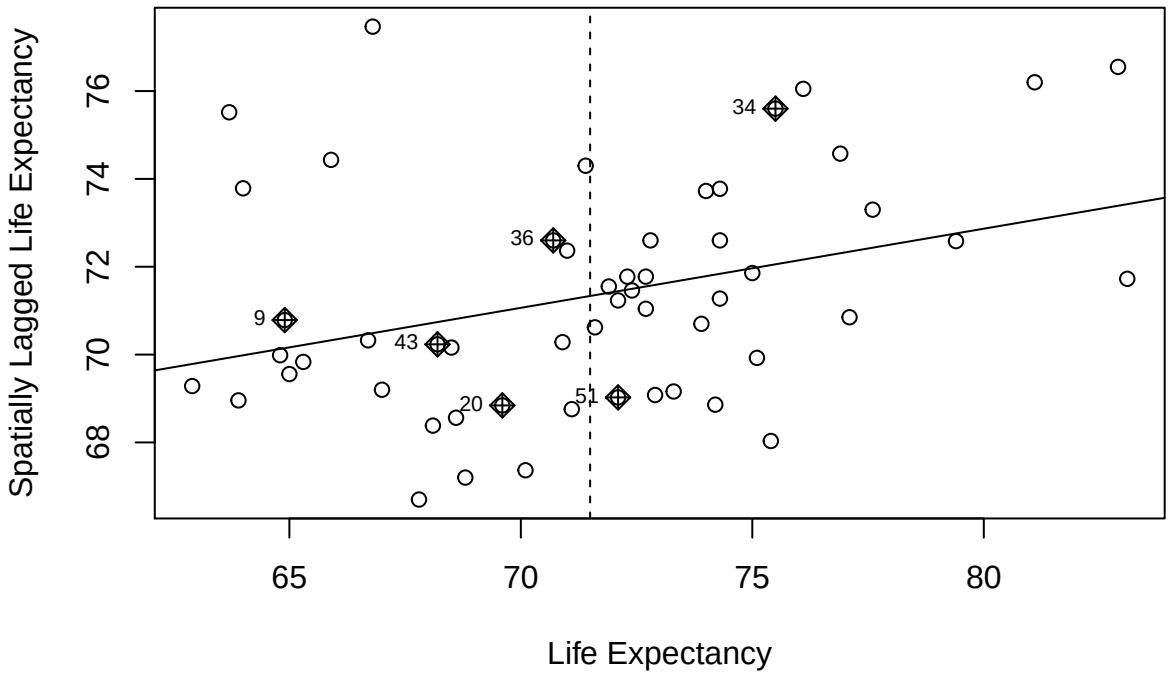
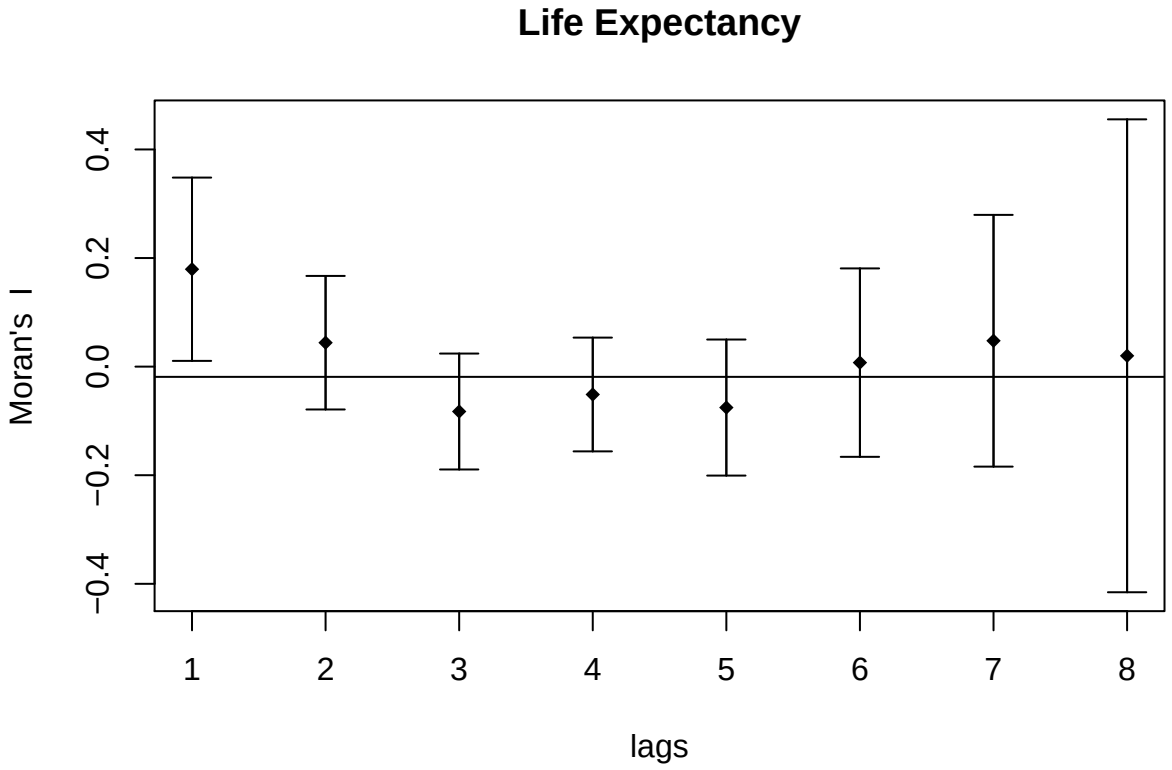


Figure 2: Spatial correlogram (life Expectancy) Spatial correlograms are great to examine patterns of spatial autocorrelation in your data or model residuals. They show how correlated are pairs of spatial observations when you increase the distance (lag) between them - they are plots of some index of autocorrelation (Moran's I)



against distance.

Model:

$$(LifeExpectancy_i = 1) = \beta_0 + \beta_1 Income_i + \beta_2 Tobacco_i + \beta_3 Poverty_i +$$

Coefficients	Estimate	P value
Income	0.00014	<0.0001
Poverty	-10.03	0.0592
Tobacco	-0.0791	<0.0001

*Significant variables: Income and Tobacco

Moran. I Results:

Observed	Expected	sd	P value
0.03114631	-0.01851852	0.01676511	0.003052587

Moran.Test Results:

Moran I statistic	Expectation	Variance	P value
0.168582233	-0.018867925	0.007149311	0.02663

For the outcome of the final model, Moran's I = 0.16 and p = 0.02663. Using that information along with the Moran Scatter Plot we observe a low/moderate autocorrelation. In addition, taking a look at the residuals to detect presence of autocorrelation helps in reviewing our model, here Moran for regression residuals was 0.1815 and p = 0.006. These results suggest a moderate positive spatial autocorrelation that is statistically significant. Adding more variables often solve the issue of autocorrelation but in this case there was no improvement upon adding few variables

Even if there is presence of autocorrelation, the final model must meet all the other assumptions of a linear model. The assumption of equal variances which often doesnt support a model that has autocorrelation, was met here. In addressing the issue of autocorrelation, two methods could have been used here: Lagging and differencing. The later used when the change in a variable from one time period to the next is hypothesized to be uncorrelated with the changes at other time points. Here we will focus on Lagging as it is more related to our data. Autocorrelation can sometimes be handled through the use of values of the target variable from the previous time period(s) as predictors in the model. A plot of the target versus the lagged target will often show a strong relationship between the two. In a multiple regression, including a lagged target variable as a potential predictor, it is just that — a potential predictor — and has to be treated as such.

Both LMerr (4.357, p-value = 0.03686) and LMlag (0.12799, p-value = 0.7205) are not significant indicating absence of spatial dependency. The robust tests help us understand what type of spatial dependence may be at work... both are non significant. This tells us we should not run a spatial lag model.

For instance running a spatial lag model on this model, Rho reflects the spatial dependence inherent in our sample data, measuring the average influence on observations by their neighboring observations. It had a positive effect (0.0026393) and but was not significant (0.72015) However, the LR test value (likelihood ratio) was not significant (0.12835), indicating that the introduction of the spatial lag term did not improve model fit. In addition, AIC was -213.9 against the original model AIC: -215.77, so it didnt improve the model, but the spatial effect went away p-value: 0.72015.

Model 2: Life Expectancy ~ Tobacco.Store.Density, Alcohol.Density,Fast.Food.Density, Corner Store, Bachelor Degree,Age_65,Median Household Income,Race:Black Model:

$$\begin{aligned}(LifeExpectancy_i = 1) = \beta_0 + \beta_1 Tobacco.Store.Density_i + \\ \beta_2 Alcohol.Density_i + \beta_3 Fast.Food.Density_i + \beta_4 Carryout.density_i + \\ \beta_5 Bachelors_i + \beta_6 Age65_i + \beta_7 Income_i + \beta_8 Race:Black_i\end{aligned}$$

Coefficients	Estimate	Pr()
Tobacco.Store.Density	-0.0998	0.00166
Alcohol.Density	-0.3108	0.04824
Fast.Food.Density	0.1452	0.13028
Carryout.density	-0.0156	0.75598
Bachelors....	12.71	<0.0001
Age..65....	12.60	0.09465
Median.Household.Income	0.000009	0.78267
Race...Black	-2.04	0.08393

Significant variables : Tobacco Store Density, Alcohol density and Bachelor Degree

By using backward selection, I utilized ANOVA F-test, to compare nested models: a "full" and a "reduced" model. It performs the Chi-square test to compare two models (i.e. it tests whether reduction in the residual sum of squares are statistically significant or not). Note that this makes sense only if the two models are nested. For interactions, I considered interactions that I thought might be important based on my domain knowledge.

That were presumably a lot fewer than the combinations among the 7 predictors. Out of the interactions I explored, I didnt see any that were significant Another observation here is that keeping some predictors poorly correlated with the dependent variable might help improve the performance of other predictors, even in the absence of interactions. Also, one effect of leaving in insignificant predictors is on p-values–they use up precious degree of freedom in small samples. But if the sample used here was large, the effect would be then negligible. As a result, the backward selection was helpful in choosing a simplified model. However, since we are looking for coefficients from the complete model, we will keep the original model. If need be, we can select the model that had a better fit later in the analysis. An important discussion to have is about whether we should control for all the variables or not. As a conclusion, It is important to keep in mind that there is no "best model", but only "useful ones".

In the above example, we saw autocorrelation in the outcome(life expectancy) this applies here as well. However for this model, the Moran's I for regression residuals is 0.09394049, and pvalue = 0.08369, indicating very minimum spatial autocorrelation in the residuals. (LMerr = 1.1671, df = 1,p-value = 0.28) and (LMlag = 0.28493, df = 1, p-value = 0.5935) We see that both simple LM tests are not significant, indicating the absence of spatial dependence The robust tests help us understand what type of spatial dependence may be at work. . . both are significant, but the lag measures is more so. The (RLMerr = 1.0577, df = 1, p-value = 0.3037) while (RLMlag = 0.17555, df = 1, p-value = 0.6752) This tells us that we should not run a spatial lag model.

Model 3: Death ~ Tobacco.Store.Density, Alcohol.Density,Fast.Food.Density ,Cornerstore density , Bachelors Degree+ Age..65 , Median.Household.Income Figure 1: Spatial correlogram (Death) Spatial correlograms are great to examine patterns of spatial autocorrelation in your data or model residuals. They show how correlated are pairs of spatial observations when you increase the distance (lag) between them - they are plots of some index of autocorrelation (Moran's I) against distance.

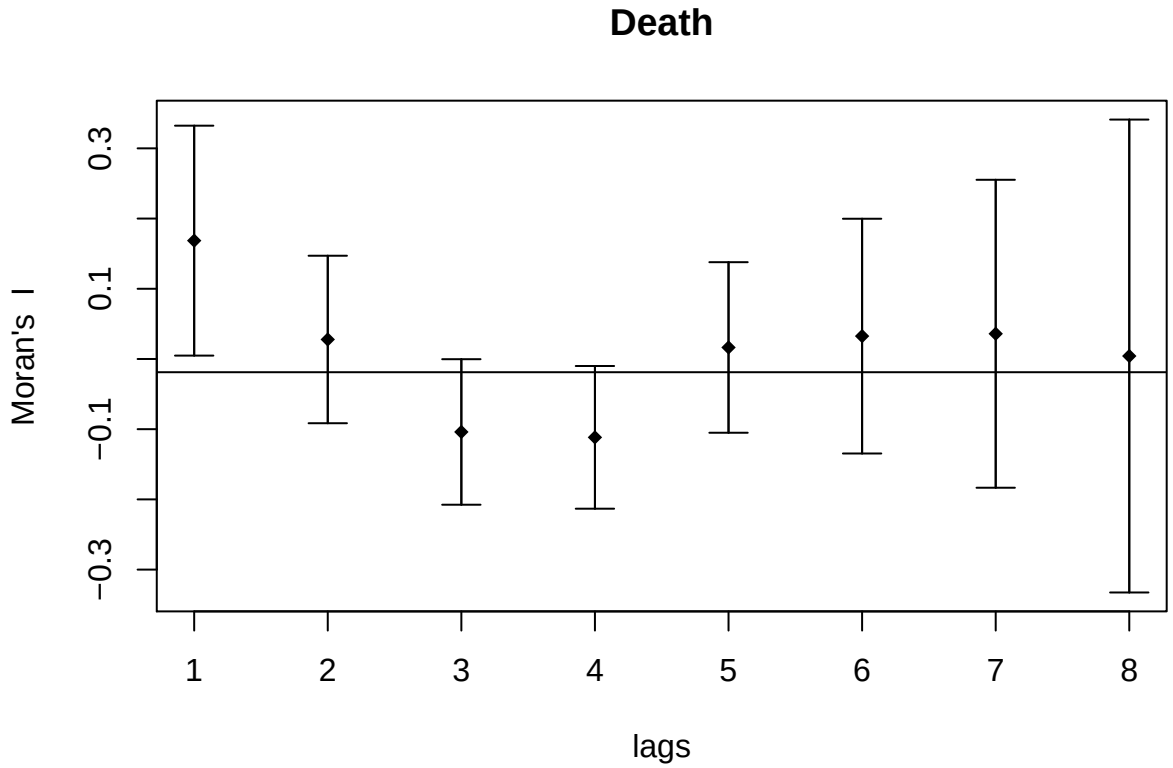
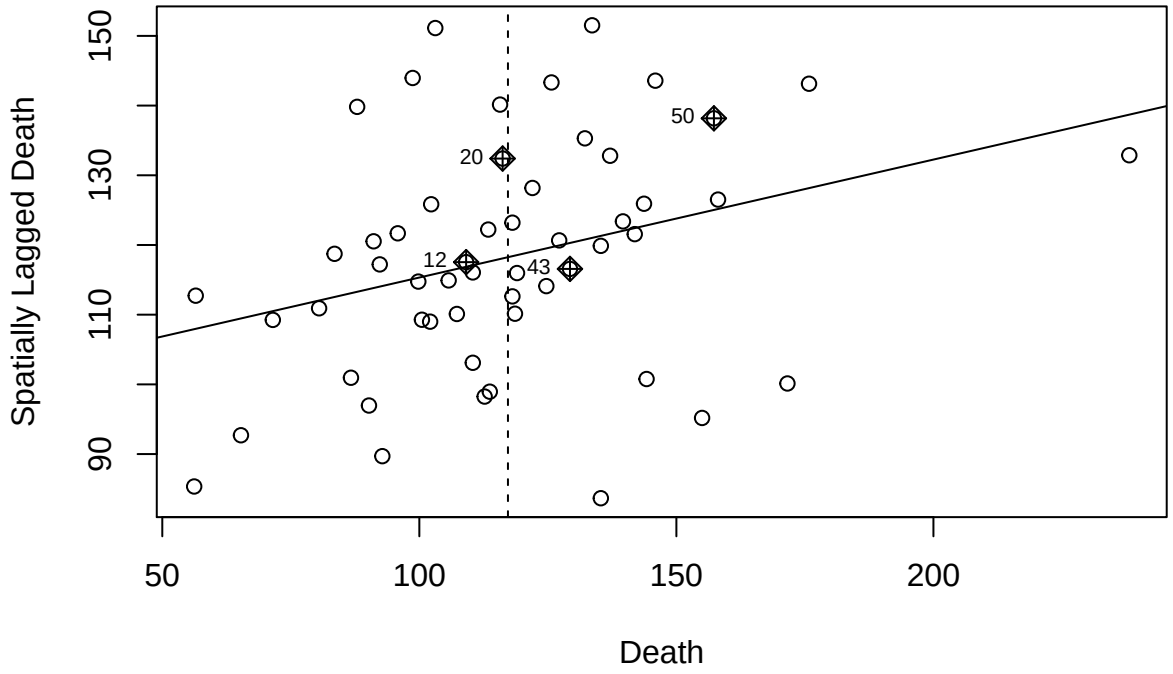


Figure 2 : Moran Test Scatterplot: The four quadrants in the graph provide a classification of four types of spatial autocorrelation: high-high (upper right), low-low (lower left), for positive spatial autocorrelation; high-low (lower right) and low-high (upper left), for negative spatial autocorrelation. The slope of the regression line is Moran's I. 2

```
## Error in `rownames<-`(`*tmp*`, value = c("0", "1", "2", "3", "4", "5", : length of 'dimnames' [1] no
```



Model:

$$(Death_i = 1) = \beta_0 + \beta_1 Tobacco.Store.Density_i + \beta_2 Alcohol.Density_i + \beta_3 Fast.Food.Density_i + \beta_4 Carryout.density_i + \beta_5 Bachelors_i + \beta_6 Age65_i + \beta_7 Income_i$$

Coefficients	Estimate	Pr()
Tobacco.Store.Density	0.519	0.023124
Alcohol.Density	2.198	0.058568
Carryout.density	0.39978861	0.293393
Fast.Food.Density	0.10469234	0.881362
Race...Black	9.29208637	0.284086
Bachelors....	-68.23339147	0.000435
Age..65....	-160.78998503	0.005090
Income	0.00009	0.699670

*Significant variables: Tobacco Density, Bachelor Degree and Age > 65

Moran. I Results:

Observed	Expected	sd	P value
0.03050022	-0.01851852	0.01628698	0.002615154

Moran.Test Results:

Moran I statistic	Expectation	Variance	P value
0.191680691	-0.018867925	0.006960602	0.01161

There is autocorrelation in the outcome "Death". However for this model, the Moran's I for regression residuals is 0.03442892, and pvalue = 0.2496, indicating very minimum spatial autocorrelation in the residuals. (LMerr = 0.15677, df = 1,p-value = 0.6922) and (LMlag = 0.59979, df = 1, p-value = 0.4387) We see that both simple LM tests are not significant, indicating the absence of spatial dependence The robust tests help us understand what type of spatial dependence may be at work... both are significant, but the lag measures is more so. The (RLMerr = 0.019645, df = 1, p-value = 0.8885) while (RLMlag = 0.46266, df = 1, p-value = 0.4964) This tells us that we don't need to run a spatial lag model.

Model 4: Years.Lost ~ Tobacco.Store.Density, Alcohol.Density, Fast.Food.Density , Cornerstore density,Bachelor Degree, Age..65...., Median.Household.Income Model:

$$(Years.Lost_i = 1) = \beta_0 + \beta_1 Tobacco.Store.Density_i + \beta_2 Alcohol.Density_i + \beta_3 Fast.Food.Density_i + \beta_4 Carryout.density_i + \beta_5 Bachelors_i + \beta_6 Age65_i + \beta_7 Income_i$$

Coefficients	Estimate	Pr()
Tobacco.Store.Density	11.15190	0.00663
Alcohol.Density	34.74101	0.09744
Fast.Food.Density	39.47315	0.00216
Carryout.density	4.81503	0.46931
Bachelors....	-1684.15565	<0.0001
Age..65....	1985.62528	0.04732
Median.Household.Income	-0.00511	0.19957

*Significant variables: Age>65 , Bachelor Degree, Tobacco Density and Fast Food Density

Moran. I Results:

Observed	Expected	sd	P value
0.01007509	-0.01851852	0.01685304	0.08976404

Moran.Test Results:

Moran I statistic	Expectation	Variance	P value
0.181159896	-0.018867925	0.007077331	0.01742

The Moran I and Moran Test indicate no autocorrelation in the variable "Years Lost".It is quite possible that the spatial distribution of feature values is the result of random spatial processes.

Model 4: death/Chronic Lower Respiratory Disease ~ Tobacco.Store.Density, Alcohol.Density, Fast.Food.Density, Cornerstore density,Bachelor Degree, Age..65...., Median.Household.Income Model:

$$(death/ChronicLowerRespiratoryDisease_i = 1) = \beta_0 + \beta_1 Tobacco.Store.Density_i + \beta_2 Alcohol.Density_i + \beta_3 Fast.Food.Density_i + \beta_4 Carryout.density_i + \beta_5 Bachelors_i + \beta_6 Age65_i + \beta_7 Income_i$$

Coefficients	Estimate	Pr()
Tobacco.Store.Density	0.0437707	0.1565
Alcohol.Density	-0.0169308	0.9157
Fast.Food.Density	-0.0064801	0.9454
Carryout.density	0.0280108	0.5862
Bachelors....	-5.3353243	0.0320
Age..65....	-16.0735031	0.0385
Median.Household.Income	0.0000873	0.0062

*Significant variables: Age>65 , Income and Bachelor Education

Moran. I Results:

Observed	Expected	sd	P value
-0.0058	-0.019	0.017	0.45

Moran.Test Results:

Moran I statistic	Expectation	Variance	P value
0.1143	-0.0189	0.0071	0.1

The Moran I and Moran Test indicate no autocorrelation in the variable "Death/Chronic Disease". It is quite possible that the spatial distribution of feature values is the result of random spatial processes.

Limitations: The presence of autocorrelation observed in some of the models was miniminal given their respective outcomes. After using the required methods to reduce autocorrelation, the results were still not very different from the initial model.