Spatial Regression_Final Paper

Geet Chawla

06/06/2020

Research Question:

The research question for this study will aim to investigate if there was any spatial autocorrelation with respect to the occurrence of Malaria in Colombia in 1988. Specifically, does the occurrence of Malaria in one municipality has any spatial correlation with the reported number of cases in neighboring municipalities in Colombia. To complement the potential correlation, the study will be controlling for other parameters like: population and precipitation levels at the municipality level.

Relevance of the Question

Occurrence of Malaria in regions around the Amazon Rainforests has always been a major development challenge in the South American sub-continent. Several high-quality research studies have also informed policy makers of valuable recommendations to curb the instance of Malaria. With emerging avenues of rich data from satellites on ecological factors, it opens up new methods to look at pressing challenges such as early remedy and preparedness for Malaria outbreaks in Colombia, or elsewhere. Evidence suggests that there is has been spatial dependence while studying Malaria within the municipalities in the Brazillian Amazon, which is similarly placed around the Amazon Rainforest as is Colombia. This makes our case stronger to vouch for indicators using a combination of ecological data for early planning and disease control. This research intends to encourage studies using satellite data that are arguable more precise, consistent, and most frequently avaiable. This also makes the case for increasing number of data points to train machine learning causal trees to predict locally-occuring epidemics.

Taking inspiration from an ongoing study at NASA to identify Malaria hotspots in the Amazon Rainforest from satellite data - this study will examine if the occurrence of Malaria has any spatial autocorrelation after controlling for ecological factors like Malaria, and other factors like population density. Overall, this research aims to encourage the shift towards technologically driven policy recommendations on locally occurring disease outbreaks.

Data and Identification Strategy

The research will use data on instance of Malaria and population per municipality in Colombia from the Center for Spatial Data Science at The University of Chicago. The data on precipitation precipitation would be used from the University of California, Santa Barbara's CHIRPS database. All data will be gathered for the year 1988. The Spatial Dependence will be diagnosed using the following regression equation using Ordinary Least Squares methods, and then corresponding Spatial models will be evaluated if a statistically significant Moran's I is observed:

$$y_{malaria} = \alpha + \beta x_{precipitation} + \beta x_{population} + \epsilon_i$$

The data for malaria would be the annual number of reported cases at the municipality level in Colombia. The precipitation, will be at the annual average level computed by CHIRPS. To gather this data, the spatial TIFs were overlayed at the municipality level (shapefile) for Colombia to capture the annual average at the municipality level.

The sample has been reduced to only the non-island geographies of the country because: after accounting for feedback from peer review - there was an important point raised that by including the islands, we are assuming that Malaria levels in municipalities on the mainland effect those on the islands the same way that they effect neighboring mainland municipalities. Since it is reasonable to think that they would not be effected in the same way, we have removed the islands for this analysis. Another reason is that the rainfall data was missing for the islands as captured by CHIRPS, and there are no cases of Malaria reported on the islands in 1988 - see **Figure ()** in Appendix. Lastly, another important reason was to evaluate the data using the most recommended Queen Matrix while performing Spatial Regressions. Intuitively, while evaluating the spillover of Malaria from one region to the other, it would not make most sense to include islands that are counted as one municipality each outside the mainland region. None of the two islands had more than one municipality each - this strengthens our decision to exclude the island geographies.

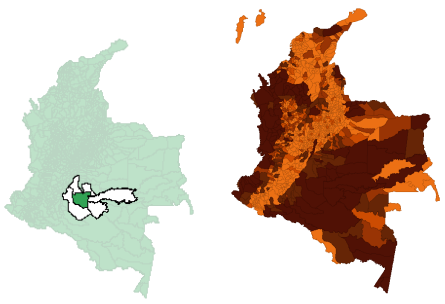

Figure 1: Queen matrix for Colombia (excluding islands); instance of Malaria in Colombia

Regression Analyses

Oridnary Lease Squares (OLS) Regression

To identify if our model even has any spatial dependence, we need to build a Spatial Weights Matrix and run a normal OLS regression. If we find any spatial dependence is when we consider running a spatial regression, otherwise we would not need to do that and a normal OLS would suffice.

Regression Output for OLS:

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :colombia_malaria_precipitation_population_noislands.dbf
Weights matrix      :        None
Dependent Variable  :    MALARI98                Number of Observations:        1066
Mean dependent var  :    174.0056                Number of Variables   :           3
S.D. dependent var  :   1105.5841                Degrees of Freedom    :        1063
R-squared           :      0.1647
Adjusted R-squared  :      0.1631
Sum squared residual:1087381151.302                F-statistic           :    104.7893
Sigma-square        : 1022936.172                Prob(F-statistic)     :   2.901e-42
S.E. of regression  :    1011.403                Log likelihood        :   -8886.840
Sigma-square ML     : 1020057.365                Akaike info criterion :   17779.681
S.E of regression ML:   1009.9789                Schwarz criterion     :   17794.596

White Standard Errors
------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT    -760.8959997     197.6080606      -3.8505312       0.0001249
                  DN       0.4315729       0.1040055       4.1495210       0.0000360
              TP1998       0.0002704       0.0001789       1.5110354       0.1310766
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER            4.441

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2     1949535.515          0.0000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                2       8828.085          0.0000
Koenker-Bassett test              2         84.048          0.0000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Moran's I (error)             0.1992        11.052          0.0000
Lagrange Multiplier (lag)          1        113.365          0.0000
Robust LM (lag)                    1          0.196          0.6578
Lagrange Multiplier (error)        1        118.921          0.0000
Robust LM (error)                  1          5.752          0.0165
Lagrange Multiplier (SARMA)        2        119.117          0.0000

================================ END OF REPORT =====================================
```

Figure 2: OLS Regression Output

In the OLS regression output we see that the R-squared is 0.16 which tells us that the model explains around 16

percent of the variation. We see that the DN variable which is the precipitation is statistically significant and the TP1998 (total population for 1998) is not significant.

Regression Diagnostics:

Multicolinearity: We observe that the model does not have substantial multicolinearity as the diagnostic is 4.4 and way less than our standard threshold of 13.

Normal Distribution: The Jarque-Bera test suggests that the data is not normally distributed. However, with the number of observations being large enough, we could think that the Central Limit Theorem would kick in, and this should not be a problem as we can use the GMM estimation methods.

Homoskedasticity: Our data is also heteroskedastic, as it rejects the null of homoskedasticity in both: the Breusch-Pagan and Koenker-Bassett test. We have hence corrected our data using the White Standard Errors.

Regression Coefficients: The coefficients of the OLS explain that with a one unit increase in precipitation, there is a 0.43 unit increase in the number of Malaria cases. Similarly, with a one unit increase in the population, there is negligible increase in the Malaria cases - this variable is also insignificant.

Spatial Dependence

Moran's I: We have a statistically significant Moran's I, which suggests that there is spatial dependence. Hence, it rejects the null of spatial randomness and suggests that there is Spatial Dependence in our model.

Lagrange Multiplier (lag): The LM test checks for the null that there is spatial randomness and there is no spatial lag - in our case we reject the null which suggests that there is some spatial lag in our model.

Lagrange Multiplier (error): The LM error test checks if there is autocorrelation in the error terms. In spatial terms, the test check if the errors are uncorrelated as the null hypothesis - we reject this hypothesis in our analysis which means that our errors are no longer uncorrelated and we have spatially correlated residuals.

Since we are rejecting the null for spatial autocorrelation in both Lag and Error versions, we need to check their robust versions to decide what model we will be going ahead with.

Robust LMs: We see that in the robust LM versions, the Spatial Lag model now turns statistically insignificant and the Spatial Lag model is still statistically significant. This indicates that we will now need to move ahead with the Spatial Error model to estimate the Spatial Autocorrelation in our model.

Spatial Error Model

When a value in one location depends on the values of its neighbors, our errors are no longer uncorrelated and may not have a normal distribution. Say we are looking at how average precipitation influences municipality-level

instances of Malaris. We will, however, expect that the way in which counties are sliced could mean that neighboring counties are more similar than distant counties, and that this similarity may influence our precipitation/Malaria estimates. This means we **violate our OLS BLUE** assumption that our error terms are uncorrelated and therefore any OLS estimates will be inefficient and biased. These spatially correlated errors are indicative of omitted (spatially correlated) covariates. Mathematically, In a normal regression we estimate a single error term, ϵ_i. In a spatial error model, we split this error into random error and spatially structured error. Here, random error (unexplained by our model) is u_i, independent identically distributed (i.i.d.) errors. Our spatially structured error is composed of the the spatial error coefficient, λ, and our original ϵ_i error term now weighted by our weight matrix. We can mathematically represent this as:

$$\epsilon_i = \lambda w_i \epsilon_i + u_i$$

We can then include this in out regression equation and run the following regression:

$$y_{malaria} = \alpha + \beta x_{precipitation} + \beta x_{population} + \lambda w_i \epsilon_i + u_i$$

Regression Output for Spatial Error Model:

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIALLY WEIGHTED LEAST SQUARES (HOM)
-----------------------------------------------------------
Data set            :colombia_malaria_precipitation_population_noislands.dbf
Weights matrix      :File:
colombia_malaria_precipitation_population_noislands_queen.gal
Dependent Variable  :    MALARI98                Number of Observations:        1066
Mean dependent var  :    174.0056                Number of Variables   :           3
S.D. dependent var  :   1105.5841                Degrees of Freedom    :        1063
Pseudo R-squared    :      0.1643
N. of iterations    :           1

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT    -726.2369328      93.9059225      -7.7336649       0.0000000
                  DN       0.4117984       0.0388833      10.5906347       0.0000000
              TP1998       0.0001676       0.0001339       1.2517790       0.2106504
              lambda       0.3180705       0.0294587      10.7971715       0.0000000
------------------------------------------------------------------------------------
================================ END OF REPORT =====================================
```

Figure 3: Regression Output: Spatial Error Model

We see that our model has a statistically significant coefficient of the $\lambda w_i \epsilon_i$ term which means that with an average of one unit increase in Malaria in the neighborhood of a unit in our geography, the instance of Malaria in the surrounded unit increases on average by .32 units.

Other Supporting Regression Models:

We also go ahead make a couple of more computations to check the robustness of our model by removing the insignificant covariate and testing for other spatial weight matrices.

Spatial Lag Model without population covariate

We test this model without the insignificant covariate population as below:

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIALLY WEIGHTED LEAST SQUARES (HOM)
-----------------------------------------------------------
Data set            :colombia_malaria_precipitation_population_noislands.dbf
Weights matrix      :File:
colombia_malaria_precipitation_population_noislands_queen.gal
Dependent Variable  :    MALARI98                Number of Observations:        1066
Mean dependent var  :    174.0056                Number of Variables   :           2
S.D. dependent var  :   1105.5841                Degrees of Freedom    :        1064
Pseudo R-squared    :      0.1619
N. of iterations    :           1

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT    -721.4410464      94.1431493      -7.6632347       0.0000000
                  DN       0.4113475       0.0390170      10.5427712       0.0000000
              lambda       0.3205859       0.0294420      10.8887364       0.0000000
------------------------------------------------------------------------------------
================================= END OF REPORT ====================================
```

Figure 4: Regression Output: Spatial Error Model without population

We see that the standard errors in our model do not change substantially at all. We would still want to keep the population estimate because it is important that we account for such an important variable while making correlations between malaria and precipitation as if there were any variation because of the population in each municipality, it would have been taken care of. None of the other coefficients also change by removing this variable. However, if we notice, our R-squared marginally drops which signals that keeping the population variable even though it is statistically insignificant makes this a more robust model.

Spatial Lag Model using KNN Matrix

Instead of using the gold-standard queen contiguity matrix, we now use K-Nearest Neighbors Matrix where we coerce each municipality to have 6 neighbors (which is also a standard practice while using KNN matrices).

We see that the coefficients remain statistically significant (apart from the population variable) and our the coefficient is rather amplified. This could be because of now the municipalities that had less than 6 neighbors now have more and hence their neighborhood becomes bigger to compute dependence on. However, the pseudo r-squared remains almost the same.

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIALLY WEIGHTED LEAST SQUARES (HOM)
---------------------------------------------------------
Data set            :colombia_malaria_precipitation_population_noislands.dbf
Weights matrix      :File:
colombia_malaria_precipitation_population_noislands_knn6.gwt
Dependent Variable  :    MALARI98                Number of Observations:        1066
Mean dependent var  :    174.0056                Number of Variables   :           3
S.D. dependent var  :   1105.5841                Degrees of Freedom    :        1063
Pseudo R-squared    :      0.1643
N. of iterations    :           1

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT    -893.0577546     116.1350675      -7.6898199       0.0000000
                  DN       0.4944901       0.0453579      10.9019658       0.0000000
              TP1998       0.0001933       0.0001313       1.4723365       0.1409300
              lambda       0.5020679       0.0454218      11.0534537       0.0000000
------------------------------------------------------------------------------------
================================ END OF REPORT =====================================
```

Figure 5: Regression Output: Spatial Error Model with KNN matrix

Spatial Lag Model using Distance Matrix

Another way to check the robustness of the study is to use a distance matrix where a neighbor is defined to be in a certain range of kilometers (or any other unit of distance) calculated from the centroid of a spatial unit from surrounding centroids of other spatial units. We took the minimum distance that would avoid having zero neighbors - this was also the default in Geoda.

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIALLY WEIGHTED LEAST SQUARES (HOM)
---------------------------------------------------------
Data set            :colombia_malaria_precipitation_population_noislands.dbf
Weights matrix      :File: colombia_malaria_precipitation_population_noislands_kms.gwt
Dependent Variable  :    MALARI98                Number of Observations:        1066
Mean dependent var  :    174.0056                Number of Variables   :           3
S.D. dependent var  :   1105.5841                Degrees of Freedom    :        1063
Pseudo R-squared    :      0.1647
N. of iterations    :           1

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT    -678.3111014     108.0799799      -6.2760106       0.0000000
                  DN       0.4161481       0.0325130      12.7994413       0.0000000
              TP1998       0.0002506       0.0001403       1.7860926       0.0740843
              lambda       0.6276556       0.0944053       6.6485203       0.0000000
------------------------------------------------------------------------------------
================================ END OF REPORT =====================================
```

Figure 6: Regression Output: Spatial Error Model with KNN matrix

By using the distance matrix, we see that the the regression output and coefficients are still statistically significant and the λ actually now contains the highest coefficient across all models and methods discussed in this paper.

Conclusion

The Moran's I in our OLS diagnostics suggests that there is spatial autocorrelation and hence rejects the null of spatial randomness. The Robust LMs motivate to go ahead with the Spatial Error model as the Robust LM (error) test rejects the null of uncorrelated residuals. The Spatial Error model evaluates the extent to which the clustering of an outcome variable not explained by the measured independent variables can be accounted for with reference to the clustering of error terms. In this sense, it captures the influence of unmeasured independent variables.

The regression outputs and the various model specifications suggest that there is spatial autocorrelation and dependence when we study the instance of Malaria in Colombia after controlling for population and precipitation using the Spatial Error Model. This suggests that the average residuals of the neighbors is correlated with the surrounded spatial unit (municipalities in our case) and rather influences the observations due to it being a neighbor. Which means that there is some exogenous variation in the residuals in our model that we needed to account for that in our model and we did that by running the spatial error model. In other words, we could think of it as a policy change or any other factor that happens in one spatial unit and that dictates this spatial dependence. Since we do not know for sure what factor it is, it is definitely to our benefit that we capture that variation caused by spatial autocorrelation in our λ term so it helps us correct the estimates of our model. λ. here shows a measure of autocorrelation of the residuals. In other words, we can see that there is moderate spatial autocorrelation in residuals because of the statistically significant coefficient.

Potential Challenges

The research may have some potential challenges in its computation:

1. Firstly, by averaging out the temperature and precipitation at the annual level, we lose out on the seasonal variation of ecological factors - we may not be able to capture season specific effects, which may be higher as compared to the average effects.

2. Since we are computing the model for the whole country, we may lose out on the potentially stronger spatial dependence observed in only the Amazon Rainforest area.

3. The model could benefit from some more covariates like surface temperature or other health and water related hygiene variables - however, due to the unavailability of gridded surface temperature data and municipality level health statistics for 1988 in Colombia - we were not able to add them into our model. Hence, we may think that these results could be slightly biased.

4. Another point to document here would be the potential changes in the computation after including the islands in the study and their precipitation data. Although trivial, it is important to note that there might be a slight change in the coefficients if we would have included the additional data.

Appendix

POLY_ID	DN	IDDANE	MALARI98	TP1998	UP1998	RP1998
1	280	44847	0	63948	4924	59024
2	-9999	88564	0	5301	1894	3407
3	463	44560	0	33798	6154	27644
4	782	44430	5	120673	104676	15997
5	1891	44001	488	95043	87958	7085
6	2968	47001	20	363350	347717	15633
7	2067	47189	10	167897	78176	89721
8	991	44378	0	7953	5778	2175
9	1016	47745	0	20671	11170	9501
10	939	8001	35	1200818	1197754	3064
11	1072	8572	2	24012	20201	14621

Figure 7: Imperfect Precipitation Obsevations for islands