Bad Neighbors and The Internet: A Geospatial Analysis of Internet Adoption

Cheng Yee Lim

May 15, 2017

3. Overview of Research Design

The research design of our paper is as follows: Section 4 describes the data used in this study. In this section, we conduct exploratory spatial data analysis of key variables and discuss the limitations of our data. Section 5 outlines the spatial econometric methodology employed in this paper. Section 5 also details robustness checks with different types of weights and the diagnostic tests we used to determine the model specification.

4. Data and Exploratory Spatial Data Analysis

Data

Data on telecommunications, institutional characteristics and country demographics have been consolidated for 195 countries in 2004, 2009 and 2014 from the World Bank and International Telecommunication Union. 2014 was selected as the most recent year of study as the dataset consists of the most complete set of relevant variables in our study. We then model the diffusion process with an evenly spaced five year interval from 2014, to 2009 and to 2004. Among the 195 countries, 119 of them were categorized as upper middle and high income countries according to the World Bank Country Classifications. Using the World Bank Country

Classifications, we defined upper middle and high income countries as developed countries and the low and lower middle income countries as developing countries.

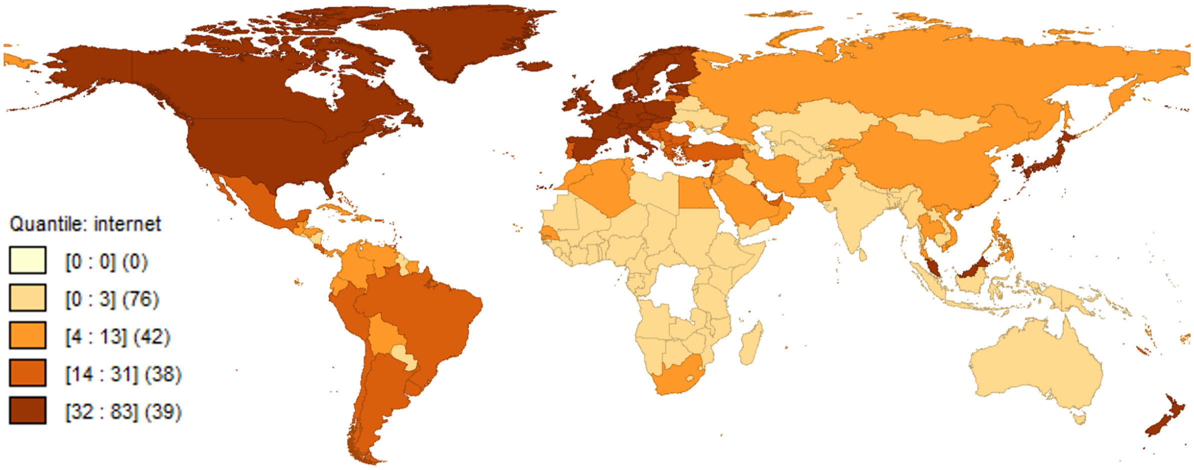
Global internet adoption has increased from 17.59% to 46.08% within a decade. Percentage of internet users, telephone fixed line subscribers and other technology use metrics per country are obtained from the International Telecommunication Union. Socioeconomic data, such as economic indicators, institution quality, and country demographics, are largely obtained from the World Bank Database of World Development Indicators. Economic capabilities are reflected by the GDP per capita of countries. Institutional quality is proxied by the number of days to set up a business in the country. Country demographics such as proportion of population living in urbanized areas, proportion of population above 65 years old, and primary education enrollment rates were also compiled in the dataset to control for demographic differences across countries. Summary statistics of the dataset by year can be found in Appendix.

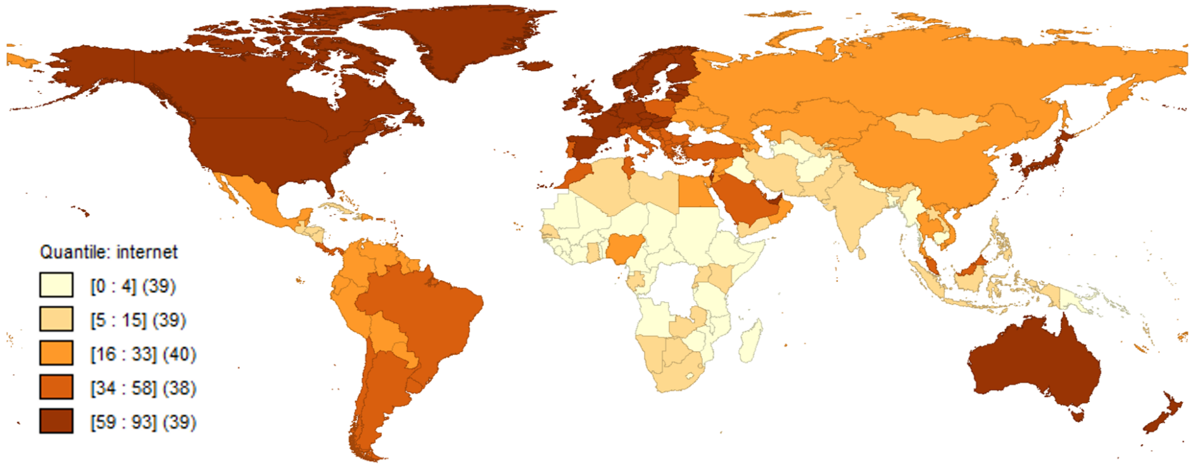
Exploratory Spatial Data Analysis

We use a series of descriptive maps to describe the spatial patterns of internet adoption in 2004, 2009 and 2014. Figure 1 and 2 illustrate the spatial distribution of internet users (per 100 people) and spatial clusters of internet adoption in the world. Figure 1 shows a series of quintile maps of internet adoption for the different time periods. From Figure 1, we can visualize that countries in closer proximity tend to be in the similar quintiles (similar shades of colors). Over time, we also observe that countries with neighboring countries in high quintiles of internet adoption tend to enter higher quintiles in the next time period. For example, the majority of countries in Western Europe and Northern Europe are part of the highest quintile in internet adoption in 2004. In the next time period (2009), we observe an change in Eastern Europe countries such as Ukraine and Belarus from 20-40% percentile to the 40-60% percentile in internet adoption. Similar trends can also be found in North Asian countries bordering Russia.

We further illustrate these spatial patterns by generating Local Moran's I cluster maps with 999 permutations. The Local Moran's I cluster maps allow us to categorize countries into statistically significant spatial clusters.

Low-low areas or cold spots are countries with a below-average internet adoption rates surrounded by similar neighbors with a below-average internet adoption rates. High-high neighborhoods or hot spots are countries with above-average internet adoption rates surrounded by neighbors with an above-average internet adoption rates. Permutations allow us to determine how likely it is to find the actual spatial distribution of a set of values by comparing them with a set of randomly generated values. The more permutations conducted, the more precise the findings. Figure 2 shows that





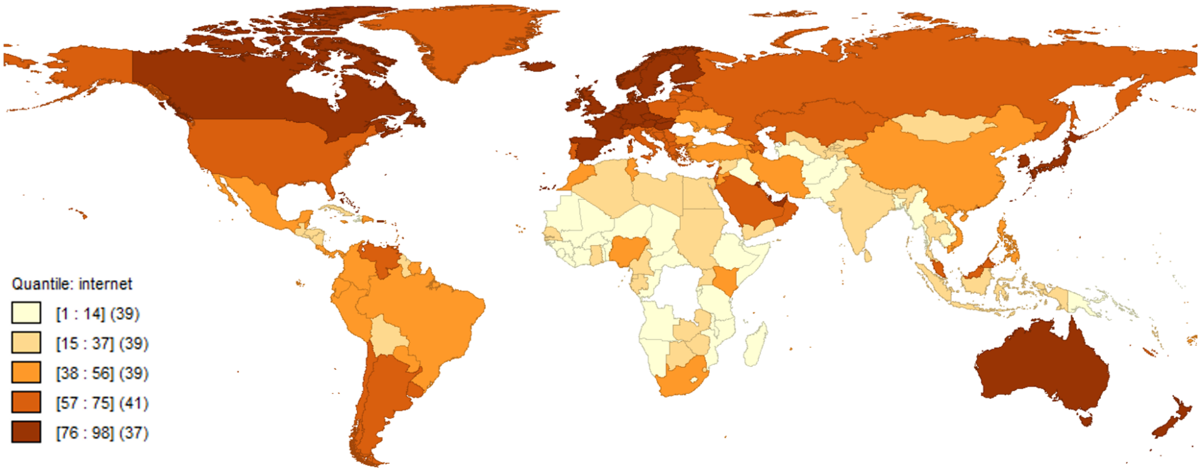
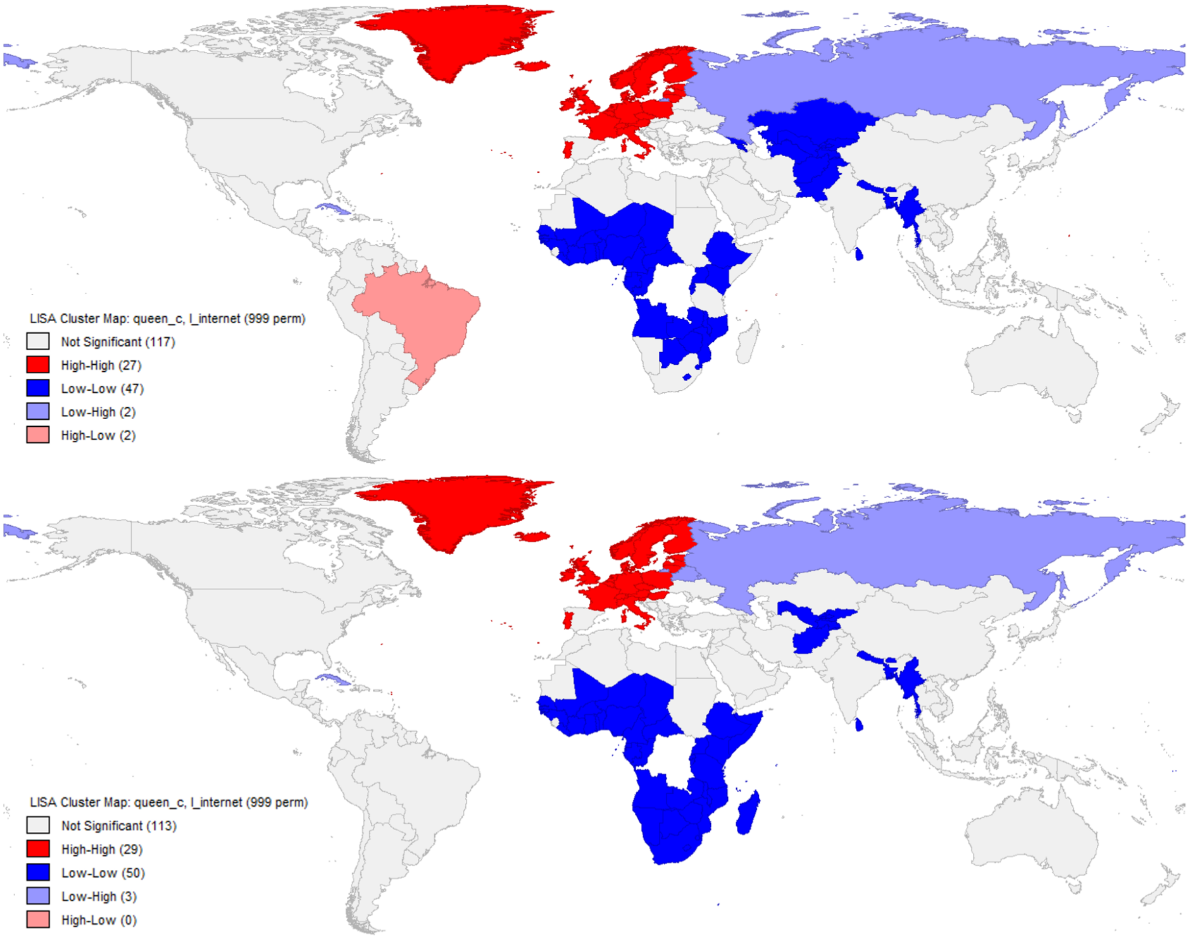


Figure 1: Quintile World Map of Internet Adoption for 2004 (Top) for 2009 (Middle) for 2014 (Bottom)



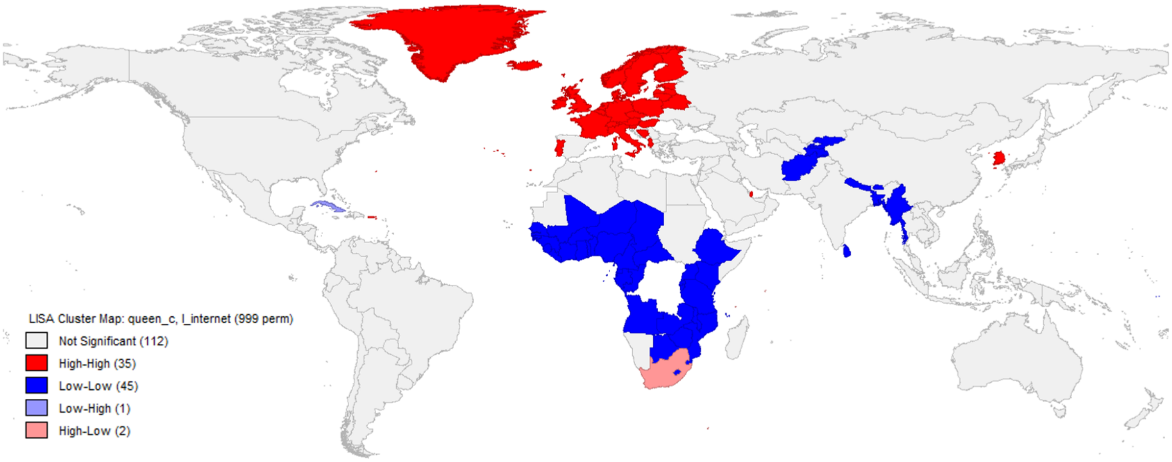


Figure 2: LISA Cluster Map of Internet Adoption in 2004 (top), 2009 (middle), and 2014 (bottom)

Cluster	Internet	GDP	Urban	Days to start a Business
Average	46.439582	11416.9252	59.03031	19.03398
Low-low	5.773879	621.6197	28.75300	25.00000
High-High	82.838954	36167.5774	74.07036	12.50000

Table 2: Descriptive Statistics of LISA Clusters

Econometric Methodology

Our dependent variable, $internet_i$, is the percentage of the population as internet users in a country. Since technology adoption is made up a series of individual adoption decisions, the most relevant measure of technology adoption is the ratio of actual to potential users. (Andres et al. 2010) This explains why the most common measures of internet adoption is the percentage of internet users in a population in the literature. The use of internet users to account for internet adoption is preferred over computer penetration rates and internet subscribers, as it includes not only internet access from the household, but also public places, such as workplaces, universities and internet cafes. (Andres et al. 2010)

According to existing literature, three main factors, economic, demographic and institutional, are pertinent in attaining widespread Internet adoption. Economic variables such as GDP per capita, level of urbanization, telephone fixed line subscribers were included as a proxy of a country's ability in providing the relevant telecommunication infrastructure in the country. Demographic characteristics, such as proportion of population aged 15-65 and above 65 and primary education enrollment rate were also included as independent variables. Lastly, the number of days taken to set up a business is used as a proxy for institutional quality in a country.

Spatial Econometric Methodology

Spatial dependence can be introduced in the regression specification in two ways. In the first specification, spatial dependence is conceptualized as an interaction between observational units. The interaction processes can occur as externalities, copy-catting, peer-effects and more. In essence, the dependent variable is defined as a function of its value at neighboring locations. The "neighbor" or spatial effect is encapsulated in the spatial lag of the dependent variable, which is represented as Wy. (S. R. Anselin A. Murray 2013)

The spatial lag at location i is thus defined as:

$$Wy_i = \sum_{j=1}^{n} W_{ij} y_j$$

where w_{ij} is the spatial weights. The convention is that spatial weights are row standardized, where $\sum_j W_{ij} = 1$. This would mean that the spatial lag, Wy, is technically obtained as a weighted average of neighboring values. This regression specification, which includes a spatially lagged dependent variable, is known as the spatial autoregressive model or spatial lag model.

The second way spatial dependence can enter the regression specification is through the error term. The unobserved effects due to neighboring locations results in non-zero off-diagonal elements in the covariance structure of random error terms. Thus, spatial error autocorrelation is a case of non-spherical error variance-covariance, where $E[\epsilon_i \epsilon_j] \neq 0$ for $i \neq j$. When identified with a spatial error term, the spatial error dependence can be specified with a spatial stochastic process model. Examples of such models are a spatial moving average form (SMA), a spatial autoregressive form (SAR) and the conditional autoregressive model (CAR). A typical spatial regression follows the specific form:

$$y = \rho Wy + X\beta + u$$

where Wy is the spatial lag of the dependent variable, X is a vector of independent variables and u, the error term. Rather than having a normal error term, the spatially dependent error term are most commonly specified:

$$u = \lambda Wu + u$$

as the SAR form, or

$$u = \lambda Wv + v$$

as the SMA form,

where Wu and Wv are the spatial lags of error term.

When cross-sectional models with spatial dependence are estimated without the spatial lags or spatial error terms, serious problems of model misspecificaiton are caused. (L. Anselin 1988) Ignoring the spatial lag of the dependent variable in a model with spatial dependence will result in an omitted relevant variable. Consequently, this will result in biased estimates and misleading inferences (L. Anselin and Arribas-Bel 2011). Thus, the methodology of spatial econometrics involves testing for the potential presence of these misspecifications and using apprropriate models to account for the spatial dependence.

Diagnostic Tests and Model Specification

The paper begins by performing an ordinary least squares (OLS) regression of the internet adoption on the independent variables. An appropriate spatial model is then selected, according to the OLS residuals and a series of diagnostics of spatial effects. The diagnostic tests consist of a series of Lagrange Multiplier tests against spatial lag dependence (LM-lag) or spatial error dependence (LM-Error) and the robust forms of those tests. The spatial models will use maximum likelihood (ML) estimation or 2SLS estimation, depending on the outcome of normality diagnostic test. If there is a strong evidence of non-normality, 2SLS estimation will be used.

Non-Spatial Diagnostics

Table presents the results of the OLS regression for internet users from 245 countries in 2015. The multicollinearity condition number is 16.24. Since the rule of thumb is to be concerned with multicollinearity condition numbers above 30, it suggests that there will not be any potential problems with multicollinearity with our specification. The Jarque-Bera test statistic for normality of errors is 22.86, which is highly significant and indicates a high degree of non-normality. Thus, 2SLS estimation will be used. Three tests, the Breusch-Pagan test, Koenker-Bassett test and the White test, were conducted to detect heteroskedasticity. The Breusch-Pagan test has a reported value of 15.93, the Koenker-Bassett test with a reported value of 9.207 and the White test with a value of 114.10. Out of the three tests, only the Koenker-Bassett test suggests that the assumption of homoskedasticity is not violated whereas the Breusch-Pagan test and the White Test suggests a strong significance and the presence of heteroskedasticity. Since both the Breusch-Pagen and White Test has such strong evidence of heteroskedasticity, we conclude that the assumption of homoskedasticity is violated.

Spatial Diagnostics

With the specification of a weights matrix, the Lagrange Multiplier statistics is calculated for the spatial lag, spatial error and higher orders of spatial autocorrelations with the OLS regression. The Lagrange Multiplier (lag) is highly significant at a 0.1% significance level, with a value of 16.8. This is further corroborated with the robust Lagrange Multiplier (lag) with a value of 10.9, which is also significant at a 0.1% significance level. The results of the spatial autocorrelation diagnostics support the presence of spatial autocorrelation.

The Lagrange Multiplier (error) test for the presence of spatial error autocorrelation. However, the rejection of the null hypothesis of LM (error) test might not only be due to the presence of spatial error autocorrelation, but also the presence of spatial autocorrelation. Thus, the robust Lagrange Multiplier (error) test attempts to disentangle this effect and only identify the presence of spatial error autocorrelation. The Lagrange Multiplier (error) test has a reported value of 6.082, thus it is weakly significant at 2% significance level. On the other hand, the robust Lagrange Multiplier (error) test has a value of 0.1372 and is highly insignificant. Thus, the statistic has moved from weakly significant to insignificant. This suggests that the Lagrange Multiplier (error) test was most plausibly picking up the error spatial autocorrelation due to the spatial lag, rather than actually having spatial error autocorrelation in its model.

Finally, the last diagnostic, Lagrange Multiplier (SARMA), is a test for higher order spatial autocorrelation. The value of the final diagnostic test is 16.95 and highly significant. The results suggest that it is probable that a higher order model specification is the correct alternative model. We have interpret the results of the Lagrange Multiplier (SARMA) with extra caution as there is two degrees of freedom in the test. The presence of two degrees of freedom could increases the likelihood of rejecting the joint null hypothesis when one marginal test has a very high value. Therefore, we would only consider higher order model if residual spatial autocorrelation is still present in the model with a single spatial parameter.

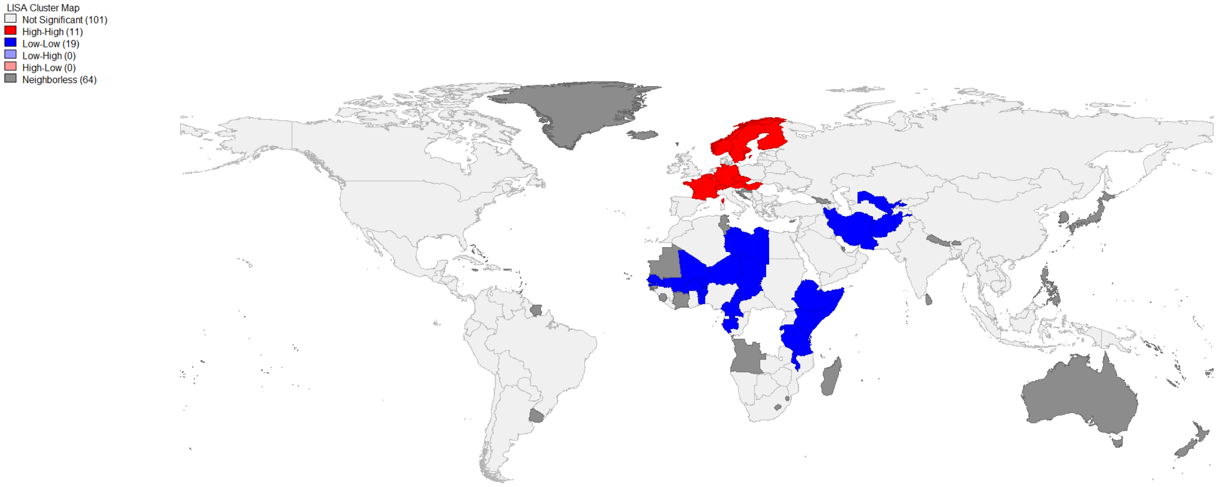
Model Specification

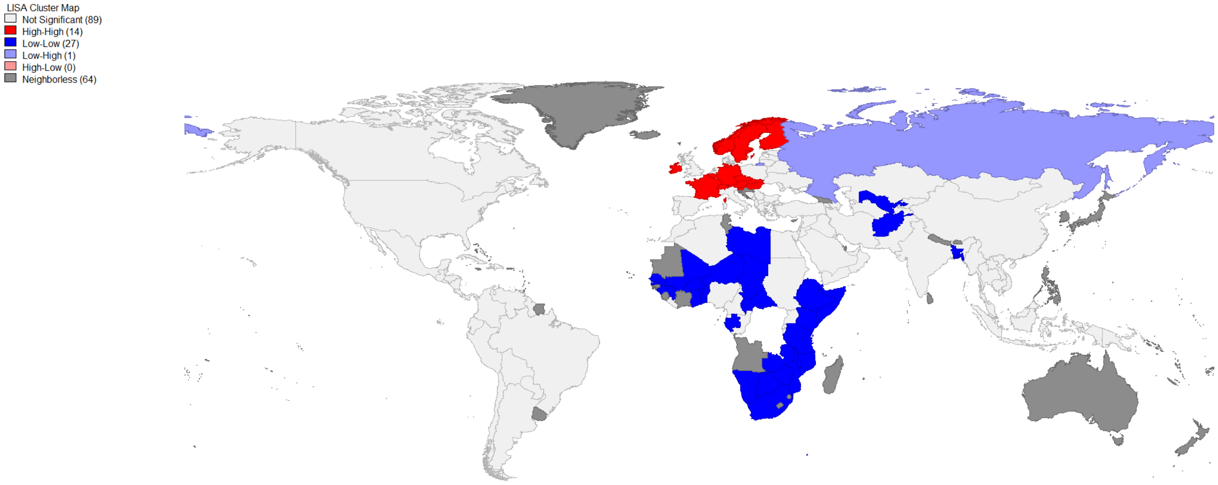
Considering the presence of spatial autocorrelation, our model is specified with a spatial lag with its independent variables from the OLS regression:

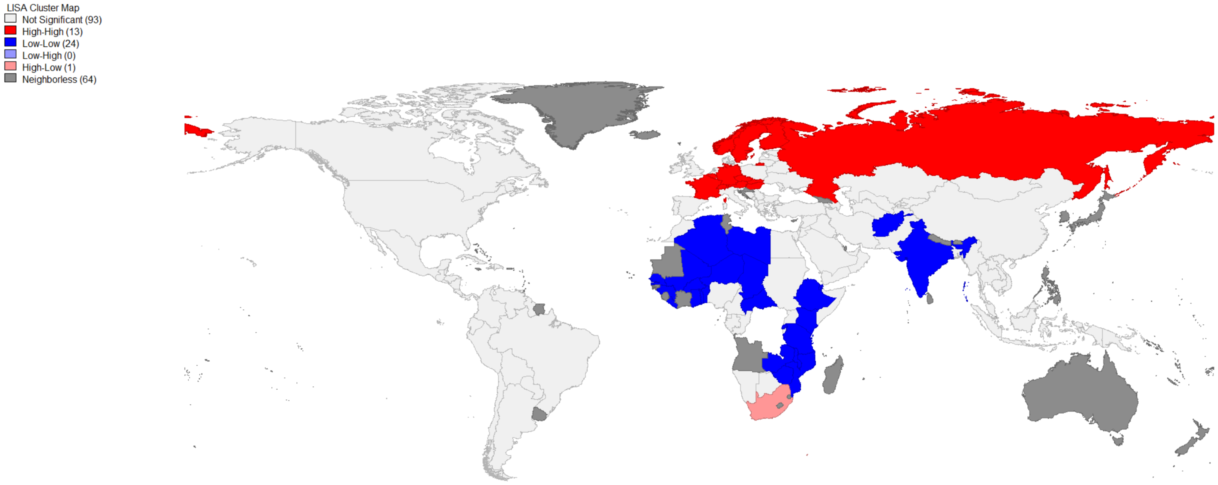
$$internet_i = \alpha + X\beta + Winternet_i + \epsilon_i$$

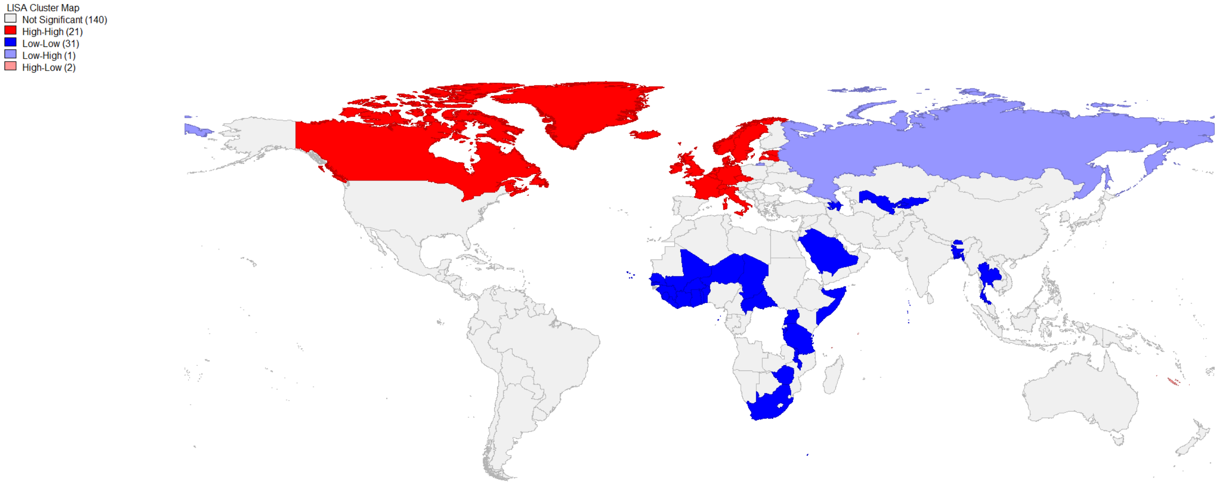
where α is a constant for all observations, X is a vector of independent variables specified earlier and β is the estimated parameter, $Winternet$ is the spatial lag of dependent variable (percentage of internet users in a country), and ϵ is the error term. Since an important assumption implicit in OLS estimation is the exogeneity of the regressors, the presence of the spatial lag of the dependent variable induces endogeneity or simultaneity. In order to correct for potential problems of endogeneity, we perform two stage least squares estimation to the model. The most appropriate instruments for the spatial lag are the spatially lagged explanatory variables.

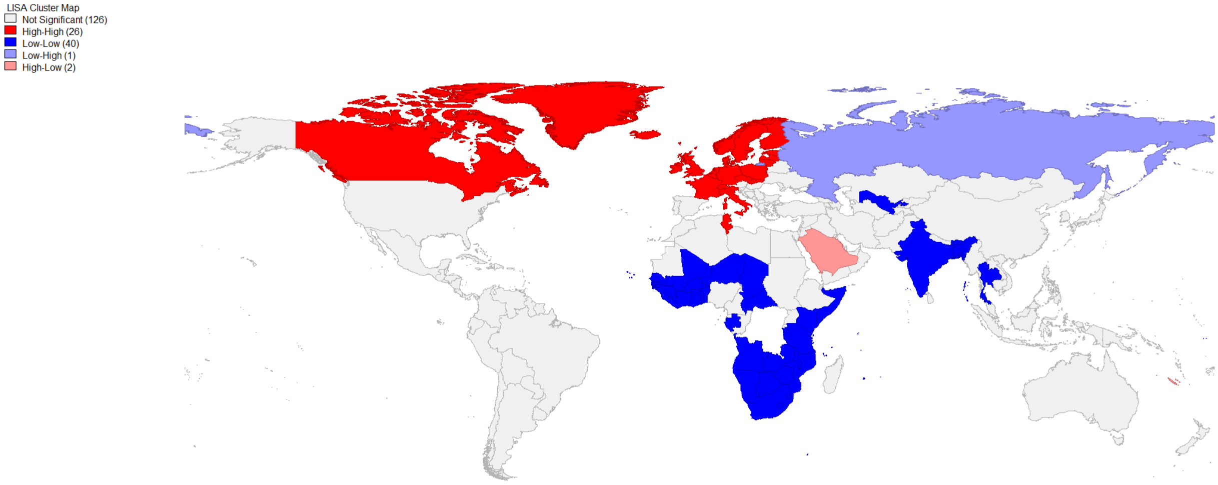
Appendix

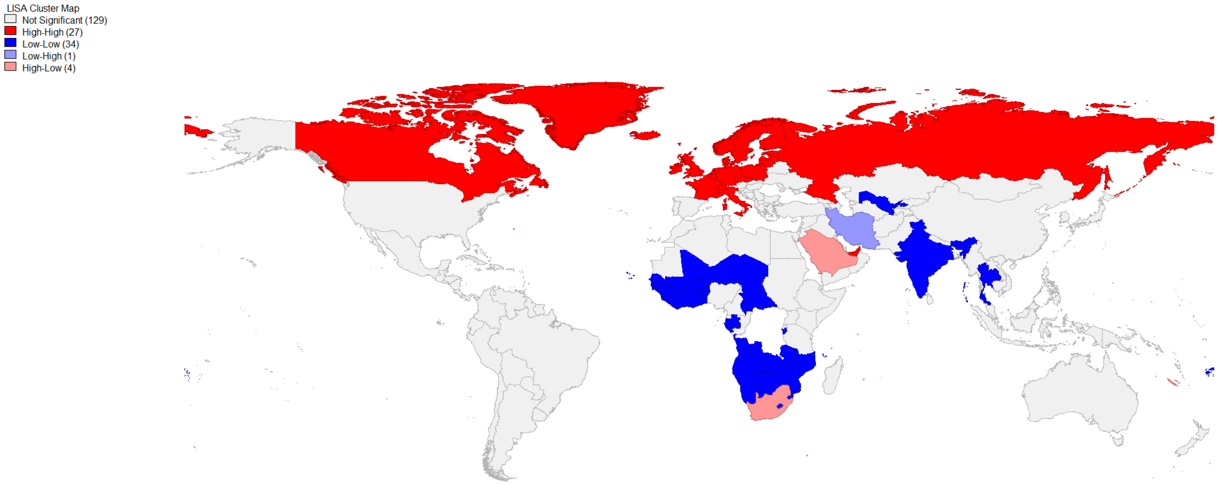

LISA Cluster Map
Not Significant (101)
High-High (11)
Low-Low (19)
Low-High (0)
High-Low (0)
Neighborless (64)

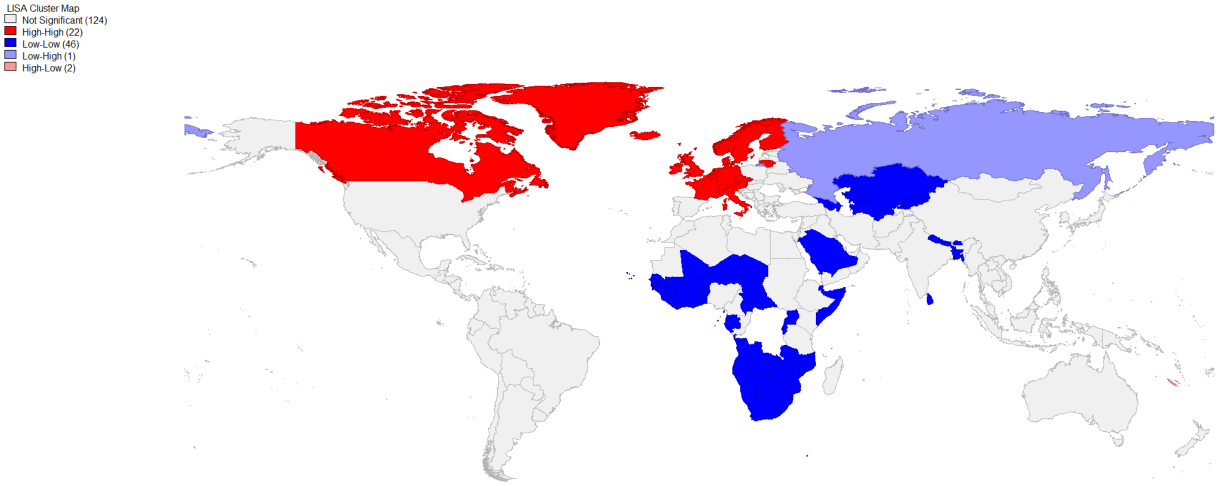

LISA Cluster Map
Not Significant (89)
High-High (14)
Low-Low (27)
Low-High (1)
High-Low (0)
Neighborless (64)

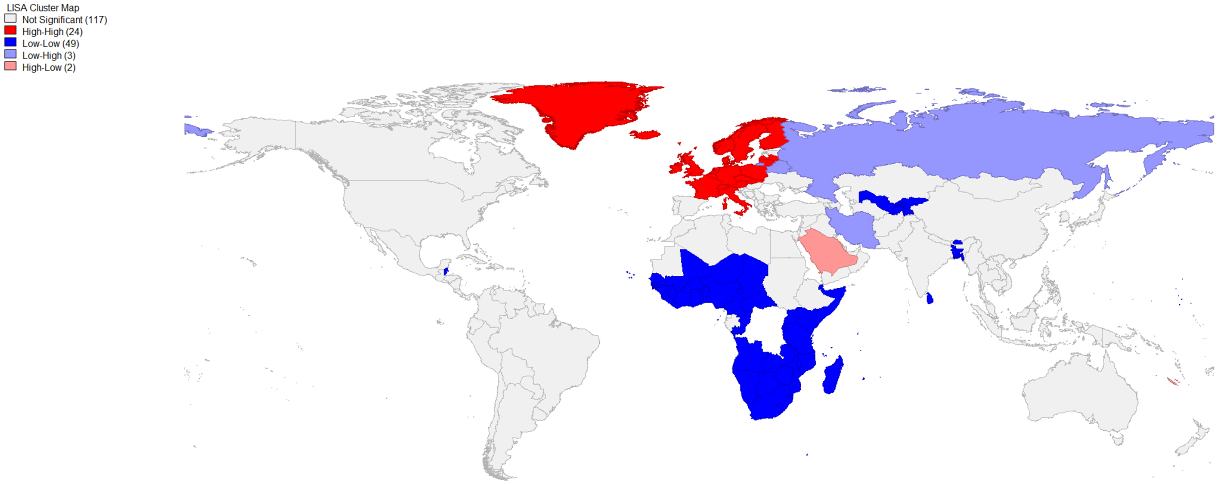

LISA Cluster Map
Not Significant (93)
High-High (13)
Low-Low (24)
Low-High (0)
High-Low (1)
Neighborless (64)

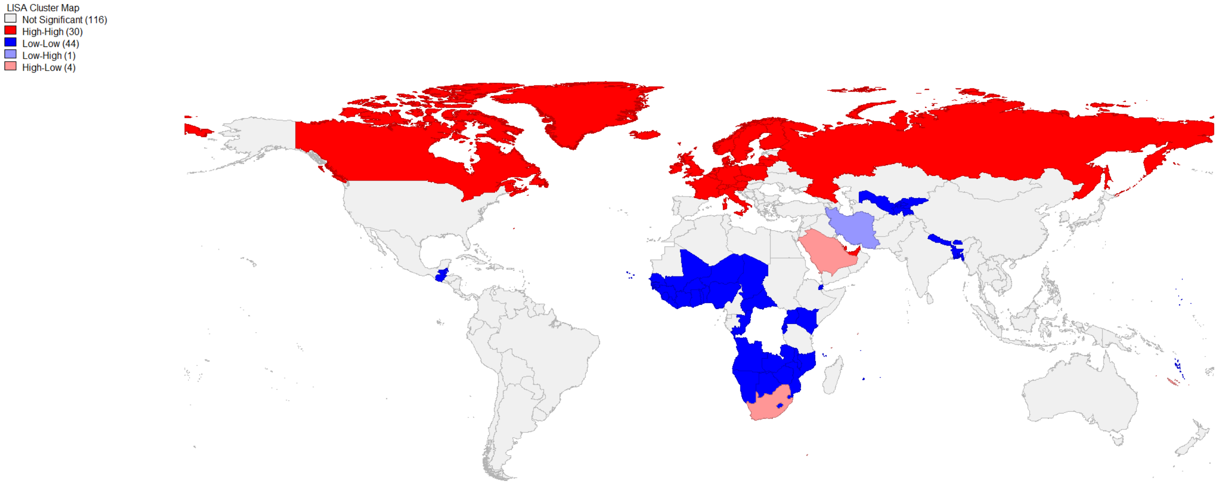

LISA Cluster Map
Not Significant (140)
High-High (21)
Low-Low (31)
Low-High (1)
High-Low (2)

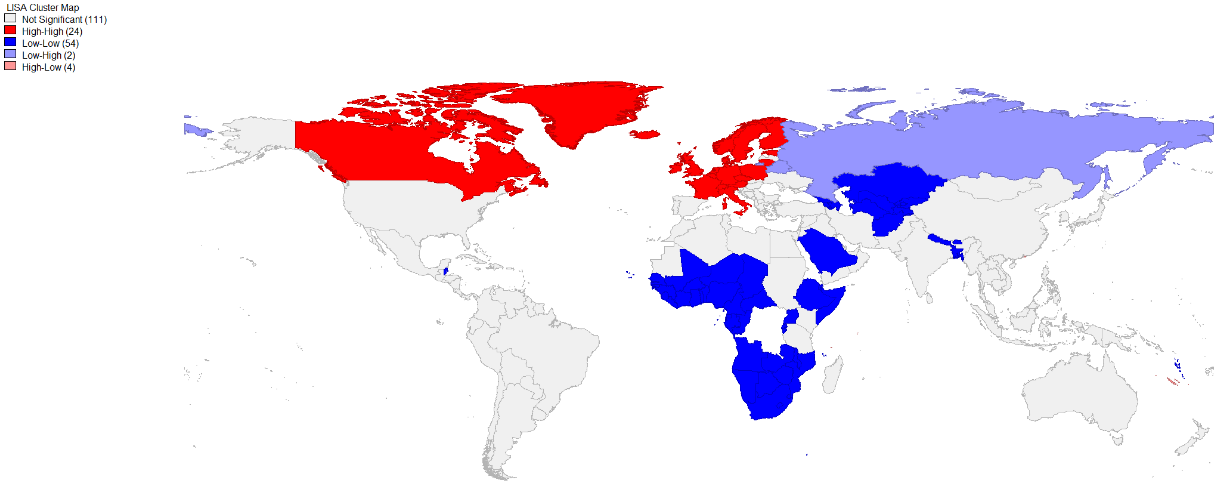

LISA Cluster Map
Not Significant (126)
High-High (26)
Low-Low (40)
Low-High (1)
High-Low (2)

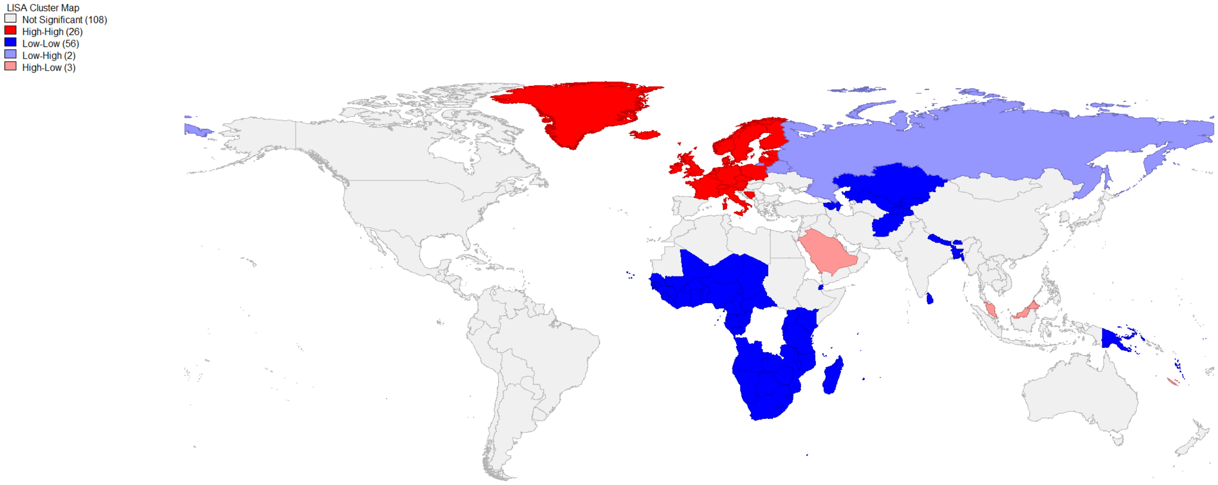

LISA Cluster Map
Not Significant (129)
High-High (27)
Low-Low (34)
Low-High (1)
High-Low (4)

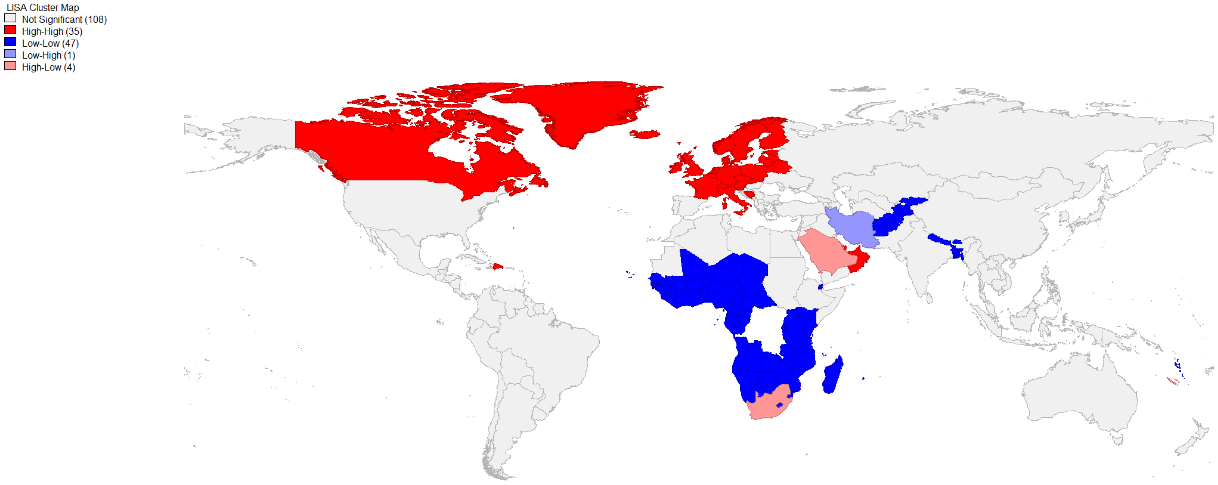

LISA Cluster Map
Not Significant (124)
High-High (22)
Low-Low (46)
Low-High (1)
High-Low (2)


LISA Cluster Map
Not Significant (117)
High-High (24)
Low-Low (49)
Low-High (3)
High-Low (2)


LISA Cluster Map
Not Significant (116)
High-High (30)
Low-Low (44)
Low-High (1)
High-Low (4)







Internet	GDP per capita	Telephone lines	Old
Min. : 1.38	Min. : 286	Min. : 0.000	Min. : 1.039

Internet	GDP per capita	Telephone lines	Old
1st Qu.:17.73	1st Qu.: 1870	1st Qu.: 3.101	1st Qu.: 3.502
Median :46.20	Median : 5484	Median : 13.356	Median : 5.884
Mean :46.08	Mean : 15829	Mean : 18.604	Mean : 8.237
3rd Qu.:70.11	3rd Qu.: 18595	3rd Qu.: 29.116	3rd Qu.:12.702
Max. :98.16	Max. :178713	Max. :132.953	Max. :25.705
NA	NA's :14	NA	NA's :10

Table 3.1: Summary Statistics of Variables in 2014

Urban	ization Days t	o Business Primar	y Education
	Min. : 8.55	Min. : 0.50	Min. : 1421
	1st Qu.: 39.31	1st Qu.: 8.10	1st Qu.: 111648
	Median : 58.66	Median : 14.00	Median : 517708
	Mean : 58.18	Mean : 22.68	Mean : 3031019
	3rd Qu.: 77.22	3rd Qu.: 28.00	3rd Qu.: 2862690
	Max. :100.00	Max. :144.00	Max. :95107120
	NA	NA's :16	NA's :66

Table 3.2: Summary Statistics of Variables in 2014

Inter	net GDP pe	r capita Teleph	one lines O	ld
	Min. : 0.22	Min. : 190.4	Min. : 0.01851	Min. : 0.6987
	1st Qu.: 7.30	1st Qu.: 1227.6	1st Qu.: 3.61227	1st Qu.: 3.4148
	Median :26.00	Median : 4576.3	Median : 16.31757	Median : 5.6907
	Mean :31.24	Mean : 14211.5	Mean : 21.12441	Mean : 7.6307
	3rd Qu.:50.80	3rd Qu.: 16772.2	3rd Qu.: 32.58993	3rd Qu.:11.3422
	Max. :93.00	Max. :152877.4	Max. :118.90455	Max. :22.2150

Inter	net GDP pe	r capita Teleph	one lines O	ld
	NA's :2	NA's :7	NA's :2	NA's :10

Table 4.1: Summary Statistics of Variables in 2009

Urban	ization Days t	o Business Primar	y Education
	Min. : 9.249	Min. : 0.50	Min. : 1837
	1st Qu.: 36.941	1st Qu.: 12.00	1st Qu.: 130205
	Median : 56.608	Median : 20.00	Median : 671683
	Mean : 56.718	Mean : 35.64	Mean : 3955244
	3rd Qu.: 75.533	3rd Qu.: 39.00	3rd Qu.: 3239838
	Max. :100.000	Max. :690.50	Max. :138368128
	NA	NA's :20	NA's :28

Table 4.2: Summary Statistics of Variables in 2009

Bibliography

Andres, Luis, David Cuberes, Mame Diouf, and Tomas Serebrisky. 2010. "The Diffusion of the Internet: A Cross-Country Analysis." *Telecommunications Policy* 34 (5). Elsevier: 323–40.

Anselin, Luc. 1988. "Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity." *Geographical Analysis* 20 (1). Wiley Online Library: 1–17.

Anselin, Luc, and Daniel Arribas-Bel. 2011. "Spatial Fixed Effects and Spatial Dependence." Citeseer.

Anselin, S. Rey, A. Murray. 2013. *The Oxford Handbook of Quantitative Methods, Volume 1: Foundations, Chapter 8: Spatial Analysis*. Oxford University Press.