Socio-Economic Determinants of Sanitation

Ridhika Agrawal
Matthew Palmeri
Mukund Kalani

Abstract: A country's access to sanitation can have widespread implications on its education levels, mortality rates and overall public health. To explore the factors that drive a country's ability to improve its access to sanitation facilities, we study the socio-economic determinants of sanitation in 112 countries for three years over a ten-year period (2005-2015). Using linear regression analysis and beta mixture models, we find that a country's information and communication technology (ICT) infrastructure, education levels, and age dependency ratio have a statistically significant impact on its access to sanitation. However, only the effect of ICT infrastructure persists over time. Using the beta mixture models, we find that there is a general movement of countries towards higher sanitation levels.

I. Introduction

Improved access to sanitation facilities is crucial to human development and survival. Generally, sanitation refers to the provision of facilities and services for the safe disposal of human urine and feces. Studies have shown that access to sanitation services determines a child's health which in turn affects the child's educational outcomes as well as the household's welfare (Dinar et al. 2016). A world bank study analyzing 70 countries found a robust negative association between access to water and sanitation technologies and both child morbidity and child mortality (Gunther and Fink 2010). Driven by the importance of access to sanitation services and its extensive impacts on communities all over the world, we aim to study the socio-economic determinants of access to sanitation services in a country.

Previous research on the determinants of access to sanitation services has been concentrated at the household level and primarily restricted to African countries. For example, a study by Mulenga et al. (2017) using linear and logistic regression analyses found that access to improved water and sanitation is concentrated among wealthier households in Zambia. Moreover, a household's wealth, gender of household head, region and type of place of residence are positively associated with access to sanitation. A study by Akpakli et al. (2018) based in Ghana found large disparities between urban and rural communities and their access to sanitation levels. This study also found significant effects of gender, age, and socio-economic status of household members on their access to sanitation facilities. Lastly, Abubakar's (2017) chi-square and ANOVA analyses revealed significant statistical differences between the type of sanitation facility households in Nigeria used and their place of residence, geopolitical zone, educational attainment and wealth. In our research, we aim to expand on these household-level studies that are limited to one continent and add to the current literature by finding the macro-level determinants of access to sanitation facilities across countries and over time.

The rest of the paper proceeds as follows - section II outlines the data used in the project. Section III talks about our empirical methodology. Section IV presents the results. Lastly, section V discusses the results and concludes.

II. Data

The data set we use is from ND-GAIN - Notre Dame Global Adaptation Initiative which primarily contains information on countries' abilities to adapt to climate change. The database's main statistic, the ND-GAIN Country Index uses two decades of data across 45 indicators to rank 181 countries annually based upon their vulnerability and their readiness to successfully adapt. Further, it contains other information relating to food availability, health care access, ecological footprints, as well as population statistics. Thus, we decided to use this data set to explore our research question.

In terms of data cleaning, we first restricted our data to the years 2005, 2010 and 2015. We chose these years since they covered a wide enough time span and due to limited availability of data. Evaluating our data further, we realized several countries had missing data for most of the predictor variables, so we dropped the following countries: Andorra, Cape Verde, Tuvalu,

Swaziland, Somalia, San Marino, Saint Kitts and Nevis, North Korea, Democratic People's Repub, Palau, Nauru, Monaco, Marshall Islands and Liechtenstein. After this process we were left with 179 countries. Next, in order to make sure there were no missing values, we forward and backward filled the values across the years, within each country (463 data points were filled). Further, since we did not want to have any missing values affecting our data, we dropped all the countries which had a missing value for all 3 years for any variable. This gave us a balanced panel data set, across 3 years, 112 countries and for 10 predictor variables. The predictor variables are defined in Appendix Table A.

III. Method

A. Linear Regression

In order to analyze the health, economic, and demographic determinants of a country's access to sanitation, we decided to model a simple linear regression. The diagnostics from an ordinary least squares model (OLS) violated the assumptions of a linear regression. Thus, we had to transform the variables, both predictor and response, to meet the assumptions.

The final regression specification for 2005 was:

$$(Access\_Improved\_Sanitation_i)^2 = \alpha_i + \beta_1 Food\_Import\_Dependency_i + \beta_2 Rural\_Population_i + \beta_3 log(ICT\_Infrastructure_i) + \beta_4 log(Education_i) + \beta_5 Dependency\_External\_Health\_Services_i + \beta_6 Water\_Dependency\_Ratio_i + \beta_7 Political\_Stability_i + \beta_8 log(Doing\_Business_i) + \beta_9 log(Age\_Dependency\_Ratio_i) + \beta_{10} Ecological\_Footprint_i + u_i$$

The final regression specification for 2010 and 2015 was:

$$(Access\_Improved\_Sanitation_i)^3 = \alpha_i + \beta_1 Food\_Import\_Dependency_i + \beta_2 Rural\_Population_i + \beta_3 log(ICT\_Infrastructure_i) + \beta_4 log(Education_i) + \beta_5 Dependency\_External\_Health\_Services_i + \beta_6 Water\_Dependency\_Ratio_i + \beta_7 Political\_Stability_i + \beta_8 log(Doing\_Business_i) + \beta_9 log(Age\_Dependency\_Ratio_i) + \beta_{10} Ecological\_Footprint_i + u_i$$

Where, $\alpha_i (i = 1,\ldots, 1)$ is the intercept for country i, $Access\_Improved\_Sanitation_i$ is the response variable measured as proportion of population in country i with access to sanitation, $\beta_1, \beta_2,\ldots, \beta_{10}$ are the coefficients for each predictor variable and u_i is the error term.

The following are the diagnostics for linear regression after the transformations:

Assumption 1: Normality of residuals

The normal Q-Q plots below show the distribution of the residuals for each of the 3 years - 2005, 2010 and 2015. Despite the presence of a few outliers on the upper-tail of each of the distributions, the residuals seem to follow a fairly normal distribution overall. We believe that the 3 linear regression models fulfill the assumptions enough to move ahead with the analysis.

Figure 1.1: 2005

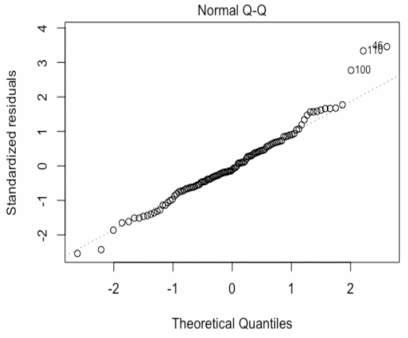


Figure 1.2: 2010

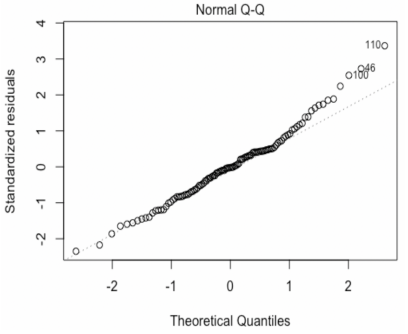
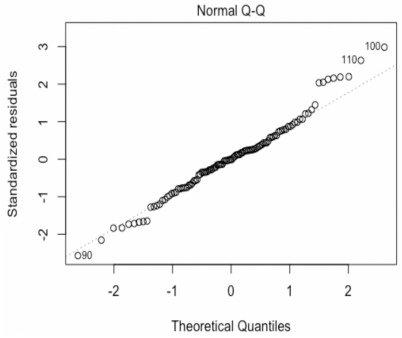


Figure 1.3: 2015



Assumption 2: Constant variance

The residuals vs. fitted plots below show that the distribution of the residuals is not completely random across the residuals = 0 line and seems to follow some pattern. This pattern occurs because of our response variable which is bounded between 0 and 100. Nonetheless, we feel that this pattern does not entirely violate the assumption of constant variance and allows us to proceed with a linear regression analysis.

Figure 2.1: 2005

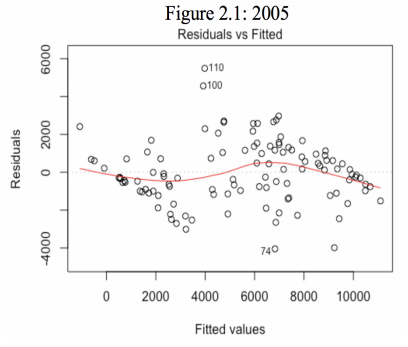


Figure 2.2: 2010

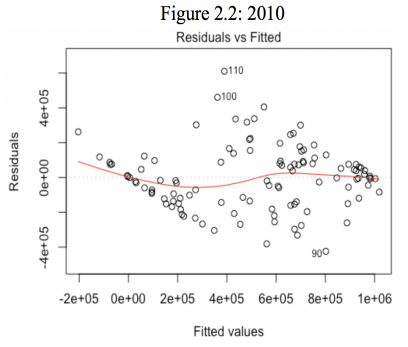
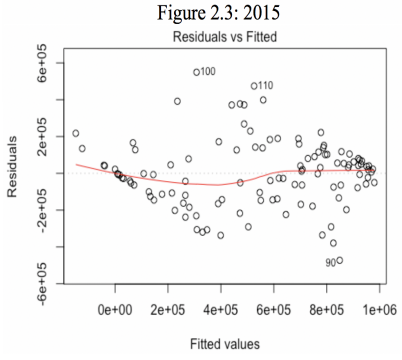


Figure 2.3: 2015



Assumption 3: Independence of error term

We conducted a durbin-watson test to check the independence of the error term. For 2005, we get a DW test-statistic of 1.88 and a *p*-value = 0.2735, for 2010 we get a DW test-statistic of 1.90 and a *p*-value = 0.328 and for 2015 we get a DW test-statistic of 1.91 and a *p*-value = 0.329. Thus, for all of these models we fail to reject the null hypothesis of independence of error term. These meet our linear regression assumption, allowing us to conduct the analysis.

Assumption 4: Outliers

Lastly, there are also some influential outliers in our data which might be biasing our results. This might be because a few of our variables have many values concentrated on the lower end (like 0). However, every country should be represented in the data and its outlier-like behavior is just a reflection of the reality of the data. Thus, we decided to keep the outliers in the data set.

Figure 3.1: 2005

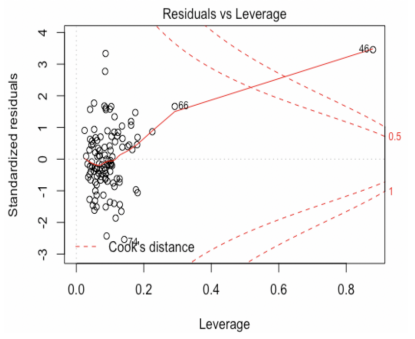


Figure 3.2: 2010

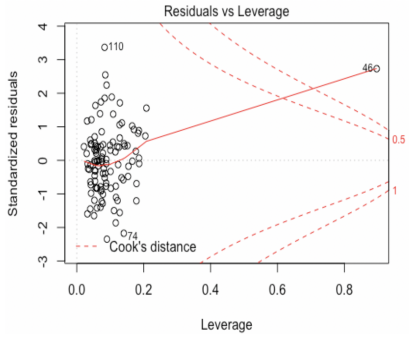
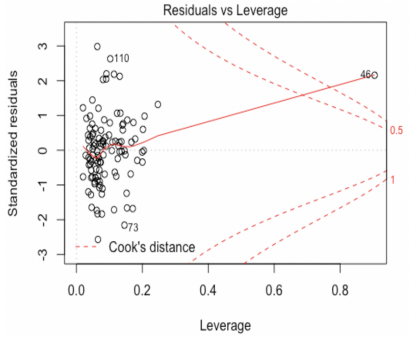


Figure 3.3: 2015



B. Beta Mixture Models

In addition to the linear regression analysis, we also used a modified version of the Expectation-Maximization (EM) algorithm to fit beta mixture models to the distribution of the Access to Improved Sanitation in 2005, 2010, and 2015. Before diving into the process, it is important to mention some of the assumptions behind mixture models, and specifically beta mixture models:

Assumption 1: The data follows the natural assumptions for fitting a beta distribution

These include mainly that the data is bounded between 0 and 1 with no density at either endpoint. With our data, this assumption is originally violated, with access to improved sanitation being on a 0-100 scale, with density at 100. This was dealt with by a simple transformation, shifting the distribution so that there was no density at 100, and then rescaling from a 0-100 scale to a 0-1 scale.

Assumption 2: The number of components needs to be chosen

This is an assumption built into any clustering algorithm and is certainly present in mixture models. With our data on sanitation facilities, it seemed appropriate from the data as well as theoretical justification to use a 2-component mixture, with one component for developing countries, and the other for developed countries.

The standard procedure for using the EM algorithm to fit mixture models is to treat the mixture parameter as the unknown quantity. The mixture parameter is the 'p' parameter in the probability density function for a two-component beta mixture model ('$Beta(x,\ a,\ b)$ is the pdf of the beta distribution for data x, and shape parameters 'a' and 'b') defined below:

$$f(x|p,\ a_1,\ b_1,\ a_2,\ b_2)\ = p\ *\ Beta(x,\ a_1,\ b_1)\ +\ (1-p)\ *\ Beta(x,\ a_2,\ b_2)$$

With the more typical fitting of gaussian mixture models, the 'E' and 'M' precede as follows:

- Expectation: Determine the proportion of the data that belongs to each component using the component parameters (a_1, b_1, a_2 and b_2 in the above pdf).
- Maximization: Generate maximum-likelihood estimates (MLE) for the component parameters using the mixture parameters generated by the Expectation step.

With the mixture of gaussian distributions, a closed form solution for the MLEs for the component parameters exists; however, with the mixture of beta distributions, this closed form approach isn't possible, due to the gamma functions in the pdf of the beta distribution. Thus, numerical methods must be performed to generate MLEs for the component parameters, which need justifiable starting points. To generate these starting points for the component parameters, we computed 'Method of Moments' estimates for each of the component parameters. Schroder and Rahmann (2017) detail the use of the Method of Moments estimates for the maximization step of fitting beta mixture models through the use of the EM algorithm; we added on to this technique, using these method of moments estimates for the starting points for numerical maximum likelihood estimates.

While not extensive, anecdotal monte carlo simulations show that this modified EM algorithm produced predicted mixture and component parameters with good coverage probabilities of the actual parameters used to generate the random data.

We used this modified version of the EM algorithm to fit two-component beta mixture models to the 'Access to Improved Sanitation Facilities' variable used as the response variable in our regression analysis above in the years 2005, 2010, and 2015.

IV. Results

A. Linear Regression

The results from the linear regression of the 2005 sample are shown below in Table 1. The results imply that out of the ten determinants of interest, only three have a statistically significant impact on a country's access to sanitation. Namely, a country's ICT infrastructure (p-value = 0.001239), education (p-value = 0.00306) and age dependency levels (p-value = 0.004765) impact its access to sanitation. The transformations in our linear model make a statistical inference less intuitive but nonetheless, it is important to take note of the determinants' direction and strength.

A country's infrastructural developments with respect to communication and broadband services imply enhanced connectivity between people and a general improvement in people's access to technology. Such developments also suggest a general shift towards a better standard of living, a big part of which is improved sanitary resources. As the results suggest, education also has a strong positive correlation with a country's access to sanitation facilities. Again, a plausible channel in this case is that with improved education levels, people have better access to job and business opportunities. These opportunities again lead to improved standards of living or a movement of people from rural to urban populations, thereby increasing their access to sanitation facilities. Lastly, we see that a country's age dependency ratio, which is defined as the proportion

of people above the age of 65 or below the age of 14, has a strong negative correlation with its access to sanitation. A plausible reason for this might be that countries with large aging populations or those with populations under the age of 14 have a relatively small working population, therefore restricting its overall economic and infrastructural development. High age dependency ratio is directly correlated to a high youth dependency ratio (populations under the age of 14), with a few exceptions. Thus, it's possible that countries with high youth dependency ratios, hence with high age dependency ratios, have high population growth which in turn reduces the amount of resources and sanitation facilities available to its people.

Table 1: 2005 Linear Regression

```
Residuals:
    Min      1Q  Median      3Q     Max
-4048.7  -998.6  -184.1  1044.2  5488.5

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                      1801.1095  2408.2063   0.748 0.456255
Food_Import_Dependency              4.8953     4.0162   1.219 0.225726
Rural_Population                  -14.1682    11.2044  -1.265 0.208955
log(ICT_Infrastructure)          3562.8686  1071.9723   3.324 0.001239 **
log(Education)                   1192.6975   318.9732   3.739 0.000306 ***
Dependency_External_Health_Services   19.7444    21.6278   0.913 0.363462
Water_Dependency_Ratio                3.6020     6.1665   0.584 0.560445
Political_Stability                -105.1279   257.9567  -0.408 0.684473
log(Doing_Business)                -290.0895   984.5728  -0.295 0.768878
log(Age_Dependency_Ratio)         -5107.7164  1769.5385  -2.886 0.004765 **
Ecological_Footprint                 -0.9351    27.7800  -0.034 0.973213
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1722 on 101 degrees of freedom
Multiple R-squared:  0.7853,    Adjusted R-squared:  0.764
F-statistic: 36.94 on 10 and 101 DF,  p-value: < 2.2e-16
```

The results from the linear regression of the 2010 sample are presented in Table 2 and show that ICT infrastructure (p-value = 0.000108) and education (p-value = 0.02778) have strong positive correlations with a country's access to sanitation facilities. However, we find that a country's age dependency ratio no longer has an impact on its access to sanitation facilities.

Table 2: 2010 Linear Regression

```
Residuals:
    Min      1Q  Median      3Q     Max
-424116 -125735   -3632   89069  610406

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                      623940.6   269964.6   2.311 0.022854 *
Food_Import_Dependency              279.6      382.9   0.730 0.466905
Rural_Population                   -352.5     1204.6  -0.293 0.770417
log(ICT_Infrastructure)          554216.7   137515.3   4.030 0.000108 ***
log(Education)                    76520.9    34275.4   2.233 0.027788 *
Dependency_External_Health_Services   -1460.0     1981.1  -0.737 0.462834
Water_Dependency_Ratio               142.7      668.7   0.213 0.831493
Political_Stability                14407.1    27310.7   0.528 0.598985
log(Doing_Business)                18598.7   105650.5   0.176 0.860615
log(Age_Dependency_Ratio)         -270366.5   185851.2  -1.455 0.148840
Ecological_Footprint                -206.6     2829.3  -0.073 0.941921
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 189400 on 101 degrees of freedom
Multiple R-squared:  0.7586,    Adjusted R-squared:  0.7347
F-statistic: 31.74 on 10 and 101 DF,  p-value: < 2.2e-16
```

Lastly, the results from the 2015 sample are shown in Table 3 and imply that only ICT infrastructure (p-value = 4.78e-06) has a strong positive impact on a country's access to sanitation. As in the previous models, education and age dependency ratio are no longer significant predictors of a country's sanitary access.

Table 3: 2015 Linear Regression

```
Residuals:
    Min      1Q  Median      3Q     Max
-472541 -114935   -1784  102045  549461

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     910899.99  316756.17   2.876  0.00492 **
Food_Import_Dependency             -19.85     384.16  -0.052  0.95889
Rural_Population                   222.06    1194.93   0.186  0.85295
log(ICT_Infrastructure)         779148.75  161154.90   4.835 4.78e-06 ***
log(Education)                   32486.69   39161.66   0.830  0.40875
Dependency_External_Health_Services  -2869.68    1959.22  -1.465  0.14611
Water_Dependency_Ratio             -70.29     661.35  -0.106  0.91557
Political_Stability              15677.25   26578.00   0.590  0.55660
log(Doing_Business)             -11070.96  126819.69  -0.087  0.93061
log(Age_Dependency_Ratio)       -240174.54  183040.39  -1.312  0.19245
Ecological_Footprint               -44.64    2855.10  -0.016  0.98756
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 190300 on 101 degrees of freedom
Multiple R-squared:  0.7556,    Adjusted R-squared:  0.7314
F-statistic: 31.23 on 10 and 101 DF,  p-value: < 2.2e-16
```

Notice that across the three years, the number of significant determinants is decreasing from three (Age Dependency Ratio, Education, ICT Infrastructure), to two (Education, ICT Infrastructure) to one (ICT Infrastructure) for 2005, 2010 and 2015, respectively. There are two narratives that can be told here, working in complement with each other.

First, there has been a general trend of an increase in world development across the years from 2005 to 2015. With the 21st century seeing economic development, education rates have been rising (Our World In Data 2020) . Similarly, the age-dependency ratio has been improving with a general decrease in world population growth rate (World Bank 2020) and better government schemes. The reason these improvements matter is because there is a threshold after which improvements or changes in a socioeconomic indicator would not make a big difference. For example, if from 2005 to 2010, the age dependency ratio decreased significantly, then a small further decrease in age dependency ratio would not have a significant impact on sanitation rates after 2010. Similarly, if from 2010 to 2015, the education rates became significantly better, then a further increase in the rates would not significantly be associated with sanitation rates. Thus, the reason the significant variables decreased across the years is because of a general trend of global development.

Next, the reason ICT Infrastructure has stayed consistently significant through the span of 10 years is once again, because of two reasons. First is due to the Digital Revolution which began in 1970 and is still underway. Thus, the progress made in ICT Infrastructure has not been saturated and has not reached its 'threshold' level. Thus, even small changes in ICT Infrastructure in countries have a huge impact on a country's overall development, and hence its access to sanitation, since several countries have a long way to go (Bell and Pavitt 1993). Second, the nature of ICT Infrastructure and technological revolution is such that even a small improvement or discovery can have major development effects, regardless of how developed a

country already is (Barro and Sala-I-martin 1997) (Jorgenson and Vu 2007). Of course, this is biased by the fact that we live in the 21st century, in the Digital Revolution.

B. Beta Mixture Models

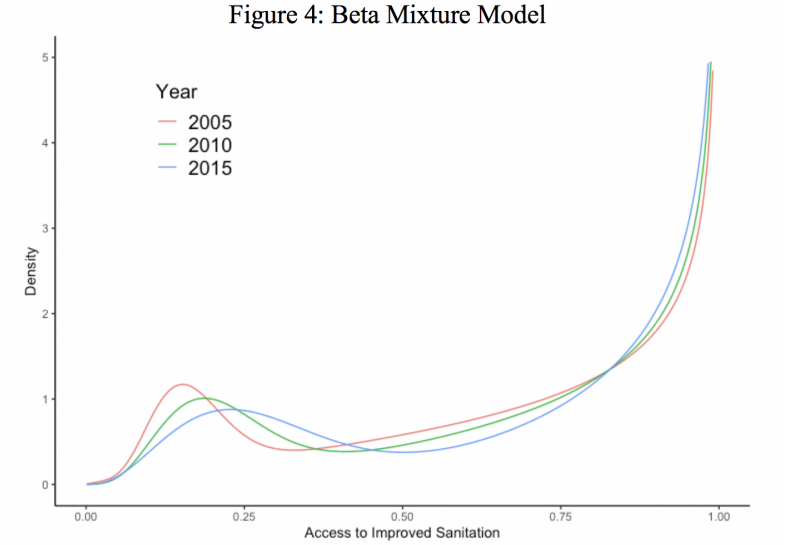
Using the EM algorithm approach outlined in the Methods section above, we fit two-component beta mixture models to the access to improved sanitation facilities variable in the years 2005, 2010, and 2015. Unlike other distributions like the Normal or the Poisson, the parameters of beta distributions don't have natural interpretations outside of the context of each other. For example, a normal distribution with a mean of 0 means that the distribution will be symmetrical and centered at 0. With a beta distribution, if one of the shape parameters is 2, the distribution could be symmetric, or it could be skewed in either direction; it depends on the value of the other shape parameter. However, in the case of mixture models, the mixture parameters have a very relevant interpretation: it represents the proportion of the data that falls into each 'category,' which in our case is country-level development.

Table 4: Beta Mixture Model Parameters

Year	Mixture Parameter	a_1	b_1	a_2	b_2
2005	0.139	6.984	35.113	1.824	0.602
2010	0.178	5.2	19.954	2.287	0.585
2015	0.238	3.774	10.702	3.416	0.619

As can be seen above in Table 4, we got mixture parameters of 0.14, 0.18, 0.23 for 2005, 2010, and 2015 respectively. But even the mixture parameters can't be interpreted outside the context of the other parameters, as just looking at these mixture parameters would lead one to believe that more countries are moving towards the developing group. However, the probability density functions of the fitted beta mixture models are graphically displayed in Figure 4 and imply otherwise. We can see a clear improvement in the lower end of the Access to Improved Sanitation spectrum. We see that the density of countries in this lower-end is moving to the right, meaning that a larger proportion of the population has access to improved sanitation facilities.

Figure 4: Beta Mixture Model



V. Discussion & Conclusion

The analysis shows that in the year 2005, age dependency ratio, education and ICT infrastructure were all significantly correlated with access to sanitation. In 2010, only education and ICT infrastructure were significantly correlated with access to sanitation, while in 2015, only ICT infrastructure was correlated with access to sanitation. Thus, the education rate of a country, the technological infrastructure advancement of a country and the proportion of the population below 14 or over 65, are all key determinants of access to sanitation. The analysis also revealed that across years, a country's socio-economic indicators such as age dependency ratio, education and hence access to sanitation, was generally improving, as can be seen by both the linear regressions and beta mixture model. Although this was not surprising, it was interesting to see these results displayed so clearly in the models across the 10 years.

To our knowledge, no other paper has done a global analysis on the determinants of access to sanitation. However, several analyses have been conducted at the country-level. Based on the intuition from those papers, the results from our paper agree with the fact that a more educated unit is correlated with higher proportion of population with access to sanitation, a higher age dependency ratio is associated with a lower access to sanitation and a more developed ICT infrastructure, and technological advancements is associated with better access to sanitation.

We believe that it would be interesting to split the data up into countries with higher access to sanitation and countries with lower access to sanitation. Researchers could then look into the socioeconomic determinants within the two groups and analyze if they differ from each other within a year, as well as compare time trends. Thus, given that this is the first paper to conduct a global analysis on the determinants of access to sanitation, there is much room for further research with different subgroups.

References

Akpakli, David Etsey, Alfred Kwesi Manyeh, Jonas Kofi Akpakli, Vida Kukula, and Margaret Gyapong. “Determinants of Access to Improved Sanitation Facilities in Rural Districts of Southern Ghana: Evidence from Dodowa Health and Demographic Surveillance Site.” *BMC Research Notes* 11, no. 1 (2018). https://doi.org/10.1186/s13104-018-3572-6.

Barro, Robert and Xavier Sala-I-martin. “Technological Diffusion, Convergence, and Growth.”*Journal of Economic Growth* 2, no.1 (1997):1–26.

Bell, Martin, and Keith Pavitt. “Technological Accumulation and Industrial Growth: Contrasts Between Developed and Developing Countries.” *Industrial and Corporate Change* 2, no. 1 (1993): 157–210. https://doi.org/10.1093/icc/2.1.157.

Gunther, Isabel, and Gunther Fink. “Water, Sanitation And Children's Health : Evidence From 172 DHS Surveys.” *Policy Research Working Papers*, February 2010. https://doi.org/10.1596/1813-9450-5275.

Jorgenson, Dale W., and Khuong Vu. “Information Technology and the World Growth Resurgence.” *German Economic Review* 8, no. 2 (January 2007): 125–45. https://doi.org/10.1111/j.1468-0475.2007.00401.x.

Mulenga, James N., Bupe B. Bwalya, and Kunda Kaliba-Chishimba. “Determinants and inequalities in access to improved water sources and sanitation among the Zambian households.” *International Journal of Development and Sustainability* 6 (2017): 746–762.

Ortiz-Correa, Javier Santiago, Moises Resende Filho, and Ariel Dinar. “Impact of Access to Water and Sanitation Services on Educational Attainment.” *Water Resources and Economics* 14 (2016): 31–43. https://doi.org/10.1016/j.wre.2015.11.002.

“Our World in Data.” Our World in Data. Accessed May 12, 2020. https://ourworldindata.org/.

Schröder, Christopher, and Sven Rahmann. "A hybrid parameter estimation algorithm for beta mixtures and applications to methylation state classification." Algorithms for Molecular Biology 12, no. 1 (2017): 21.

University of Notre Dame. “Home // Notre Dame Global Adaptation Initiative // University of Notre Dame.” Notre Dame Global Adaptation Initiative. Accessed May 12, 2020. https://gain.nd.edu/.

“World Bank Group - International Development, Poverty, & Sustainability.” World Bank. Accessed May 12, 2020. https://www.worldbank.org/.

Appendix

Table A.

Variable	Definition
Access to improved sanitation facilities	Proportion of population with access to excreta disposal facilities
Food import dependency	Proportion of cereal consumption obtained from imports
Rural Population	Proportion of population living in rural areas
ICT Infrastructure	Average over the scores of the four sub-indicators - mobile subscription per 100 persons, fixed phone subscription per 100 persons, broad-band subscription per 100 persons, and percent of individuals using internet
Education	Proportion of tertiary-education aged population enrolled in tertiary education
Dependency on external resources for health services	Percentage of external resources (e.g. bilateral payments, NGO operations) in the total national health expenditure
Water Dependency Ratio	Proportion of total renewable water resources originated outside the country
Political Stability and non-Violence	Perceptions of the likelihood of political instability
Doing Business	Average of rankings of 40 indicators as reported by the World Bank Doing Business Index for each country
Age dependency Ratio	Proportion of population above 65 or below 14
Ecological footprint	The number of hectares of land and water, both within and outside the country, that are needed to meet the average demand on ecosystems services