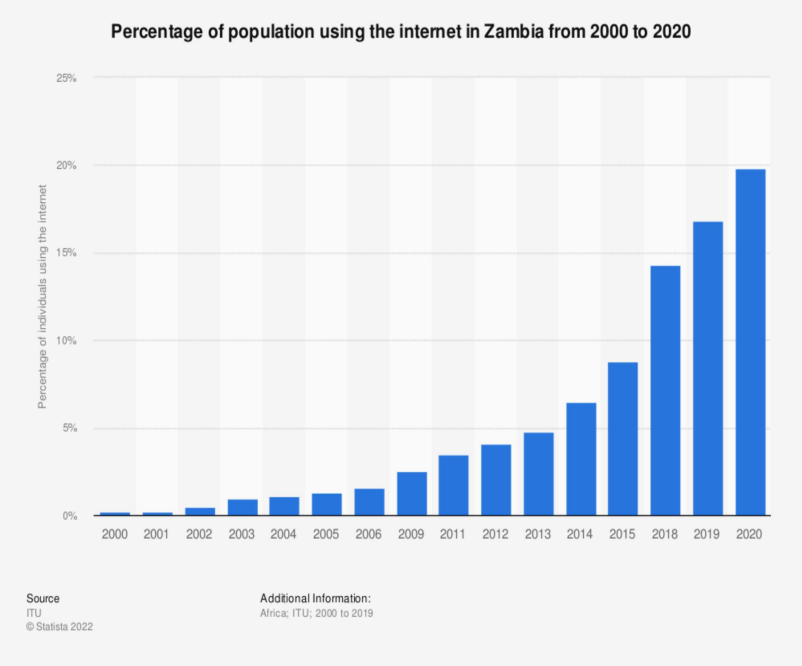
Exploring the Correlation Patterns of Internet Connection and Election Outcomes in Zambia

Bernardi Marta

Introduction to Data Analysis with R - Final Project - Winter 2023 - CEU

Executive summary

Zambia has been characterized by an increased internet access across the early '00s as shown by the graph, taken from Statista using elaborations from the International Telecommunication Union (ITU) reported below:



- As it is visible from the histogram around 2016 there has been a **significant increase in the internet access** preceded by a steady increasing trend across all the previous years, and the literature reports this spike to be related mainly to a more diffused access to mobile internet. At the same time Zambia has hold two main elections in the period around the increase in the mobile connection bandwidth: one in 2011 and one in 2016.
- Therefore, the aim of this project is to assess how the possibility to connect to the internet might affect socio-economic structures of political governance through channels such as: a change in social cohesion, an effect on democratic participation and open competition or also a change in the political game that reshuffle power dynamics across parties thanks to a change in instrument for political communication.

Many are the potential channels through which internet connection could affect outcomes related to political governance and more broadly the relationship of citizens with the state.

- The specific goal of the report is to **assess whether heterogeneity in mobile access connection is correlated with the presence of different levels of voters turnout, that is proxying electoral participation, or with different levels of valid votes, proxying a better understanding of the democratic competition rules**.

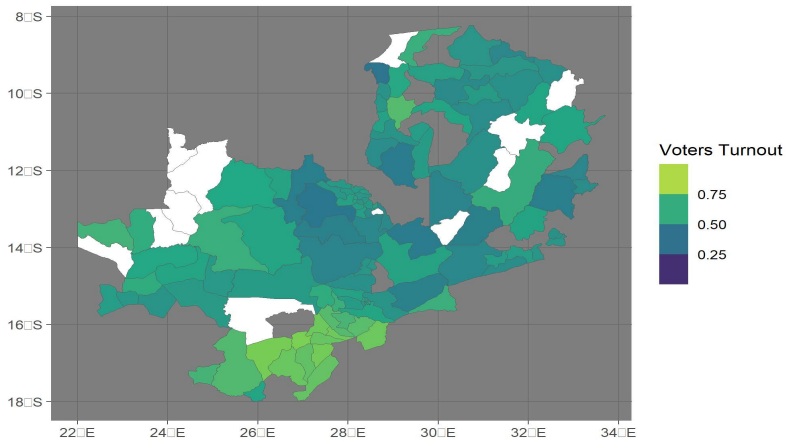
Three main datasets are used:

Firstly from the open election archives the shape of the electoral district of Zambia is taken and putted together with data coming from the OpenCellID website with an API to obtain the allocation of mobile phone towers to the electoral districts.

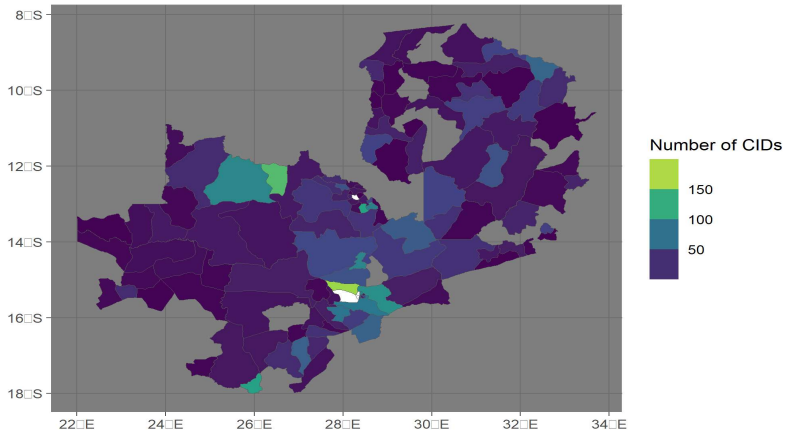
Then always from the open election archives electoral district level data on the electoral outcomes of the 2016 lower chamber election are employed.

The the two following figures is visible how both voters turnout and the number of mobile cells towers have a different geographical distribution across districts:

Geographical Distribution of the voters' turnout



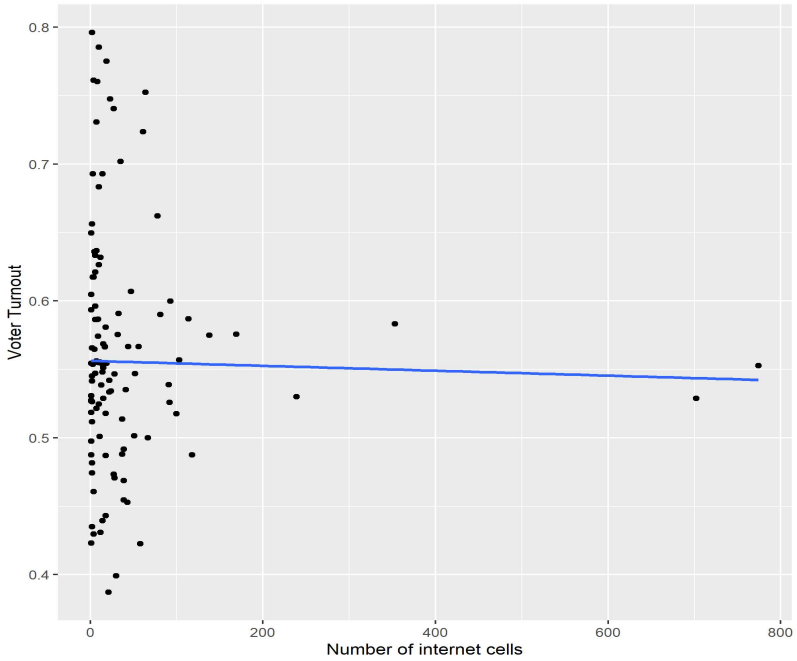
Geographical Distribution of internet mobile cells



The research project finds that there is **little to no correlation between both the electoral participation, the number of eligible votes and the internet access** as visible in the correlation plot below.

This could be happening because there is a lot of **measurement error** in the mobile access given that the data are taken from an open source system where users voluntarily register the presence of internet cells.

Or alternatively it could e because there is already a minimum level of internet everywhere by 2016 and therefore it is no more so crucial for electoral outcomes.



The project also explores the patterns of partisan vote across geographical units, finding a different geographical division for the votes to the two main parties with NRP taking the majority of its votes from the North-East and MMP winning across the South West.

The materials needed to replicated the project can be found at : https://github.com/M0arta/R_Project_Marta.git (https://github.com/M0arta/R_Project_Marta.git)

Introduction:

The project employees 3 dataset with the aim to combine geo-referenced election outcomes at electoral district level in Zambia with OpenCellID data coming from a public access API containing information on the presence of active internet cells and their coverage.

Firstly, I upload the data containing the geo-referenced information on internet access downloaded from : https://opencellid.org/downloads.php (https://opencellid.org/downloads.php) and I rename the columns so that it is understandable which values is what and I keep only interesting columns , to obtain the data as below as save them in a csv.

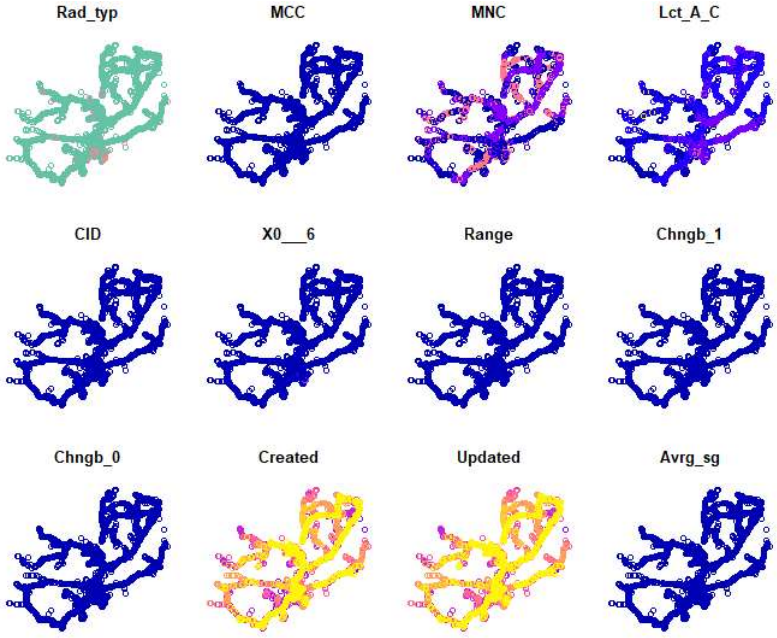
Mobile access in Zambia raw data

CID	Location_Area_Code	lat	lon	Range	Created
62893	10004	28.27491	-15.47154	2822	1459772900
2972	11108	28.33001	-15.43928	6474	1459774040
62892	10004	28.28792	-15.47047	1300	1459772900
33083	1	28.25304	-15.46148	1000	1369370001
33081	1	28.25134	-15.45296	1000	1369353306
33933	1	28.24974	-15.44518	1000	1369353496

Then I transform the data from csv to a shape file creating a variable called geometry containing the information now stored in the variables latitude and longitude.

I keep only variables for which there are no missing data for longitude and latitude, otherwise I do not know where to place them in the map and I assign a coordinate reference system so that I can compare this map with the map that I will use later containing the data on the elections, to obtain the shape file of the mobile data as below:

The data on mobile access are now in shape-file format and look like this :



Here we can already see the shape of Zambia even if the borders are not clear, we see how the different variables in the dataset are distributed across the coordinates of Zambia. For now this is not particularly indicative of anything, beside that the internet access data have been successfully converted into a shape-file.

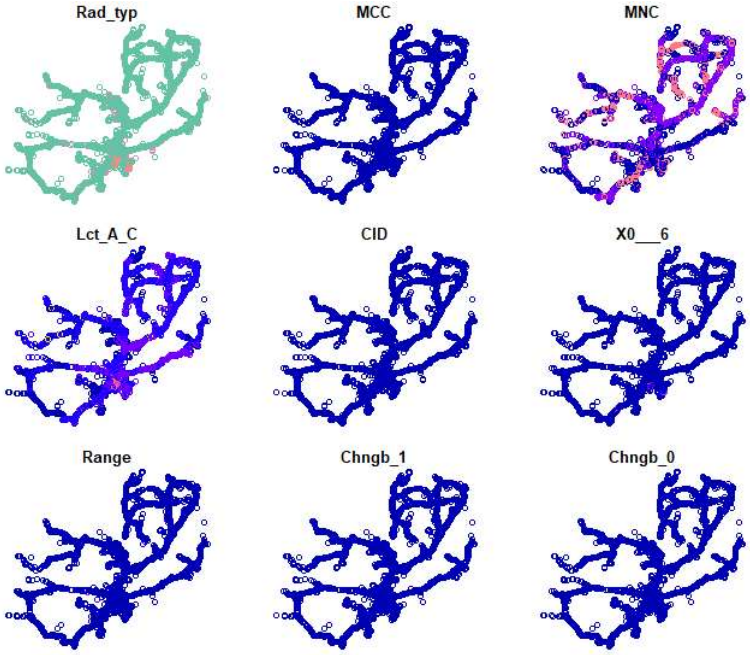
Now to give Zambia a shape and to upload the remaining part of the data necessary for the analysis I open from https://electiondataarchive.org/data-and-documentation/georeferenced-electoral-districts-datasets/ (https://electiondataarchive.org/data-and-documentation/georeferenced-electoral-districts-datasets/) the shapes of the Zambian electoral districts, I assign the same coordinate reference system that I assigned to the mobile data above and I obtain the data as below:

```
## Reading layer `gred' from data source
##   `C:\Users\Bernardi_Marta\Downloads\R_Project\gred.shp' using driver `ESRI Shapefile'
## Simple feature collection with 150 features and 5 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: 21.99937 ymin: -18.07947 xmax: 33.7057 ymax: -8.22436
## CRS:           NA
```

Shape of electoral districts in Zambia

ctr_n	ctr	yr	cst_n	cst	geometry
Zambia	894	2006	Mfuwe	94	POLYGON ((32.23695 -11.2491…
Zambia	894	2006	Kawambwa	61	POLYGON ((29.34078 -9.27867…
Zambia	894	2006	Nyimba	128	POLYGON ((31.42754 -14.0779…
Zambia	894	2006	Solwezi West	146	POLYGON ((26.15134 -11.9382…
Zambia	894	2006	Kapiri Mposhi	53	POLYGON ((28.38889 -13.8794…
Zambia	894	2006	Lufwanyama	72	POLYGON ((27.28778 -12.3438…

Plotting the data we can see how they report the coordinates of each electoral district in Zambia:



Data exploration

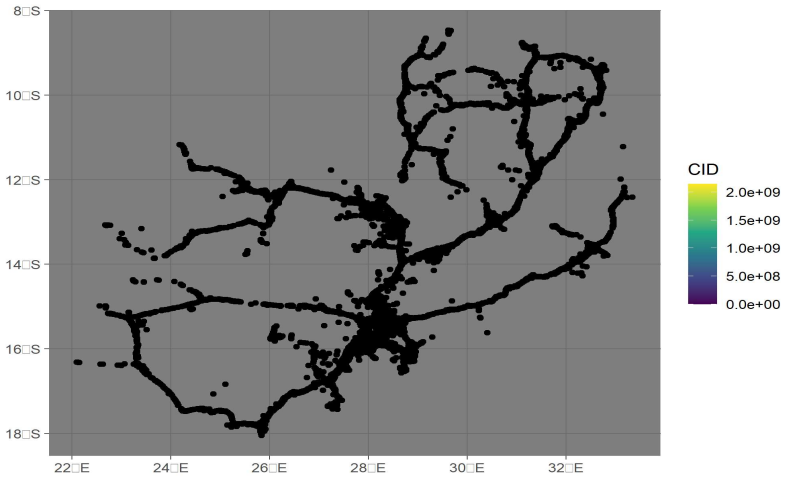
Now that I have both the mobile phone information and the shape of the electoral districts I can start exploring the variables of interest to answer to my research questions:

1) Which is the effect of internet access, measured as the presence of an active registered CID in the electoral district, on electoral outcome, partisan vote and number of valid votes?

2) How the results change when the presence of internet access is measured using the fact that there is at least one cell in the district ?

I start by looking at how the presence of internet cells are distributed in the geography looking at all the data available regardless of the year:

Presence of Internet Cells



It is visible how considering all the years at the same time there are CID bringing internet all across the country.

Now to consider only data for the years of interest for the elections I slice the dataset into years using the column called created.

To do so I first convert the timestamp into a date, it is a Unix timestamp so it is indicating the number of seconds from a fixed date and is unreadable as it is.

The I create a list of years and I loop through it to create one dataset for each year. I'm interested in years 2011 and 2016 that can be seen below:

Mobile access in 2011

Radio_type	MCC	MNC	Location_Area_Code	CID	0	lat	lon	Range	Changeable=1	Changeable=0	Created	Updated	Averag
UMTS	645	2	115	95744	0	32.76329	-9.330826	1000	3	1	2011-05-07 11:45:38	1305121951	
UMTS	645	1	1008	103838	0	28.35777	-15.371933	1000	1	1	2011-06-07 13:18:51	1307452731	

Mobile access in 2016

As it is visible the data for 2011 contain only 2 observations so the rest of the analysis will be focused on the 2016 elections.

I now open the mobile data for 2016 as a sf object (a way to call the shape files when openend with specific libraries in R) and I assign again always the same coordinate reference system.

Then I check and compare the coordinate reference system and make sure they match.

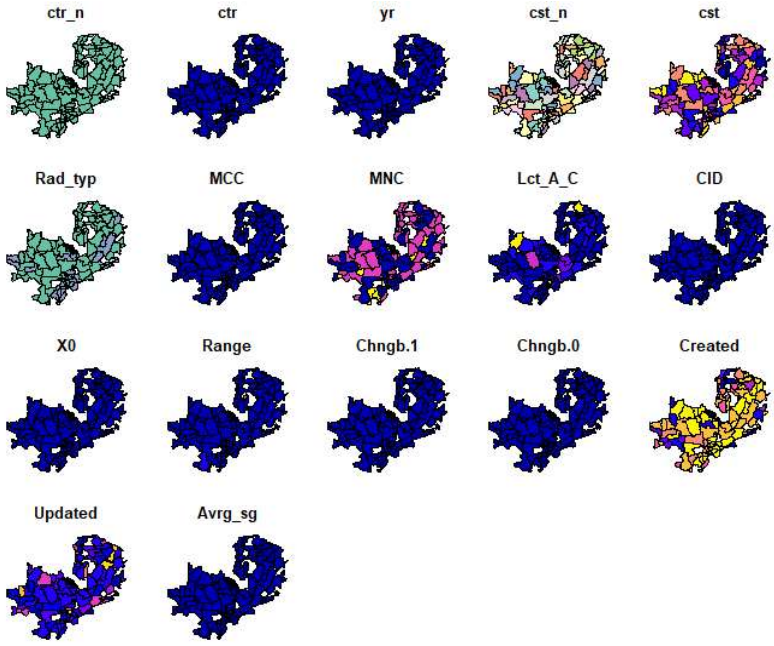
At this point I can merge the file with the shape of the electoral districts with the one of the mobile access in the year of interest for which the data are available and then write the merged dataset into a shapefile, so that I can see how many cells were in each district.

And the data look as below:

```
## Reading layer `y2016' from data source
##   `C:\Users\Bernardi_Marta\Downloads\R_Project\y2016.shp' using driver `ESRI Shapefile'
## Simple feature collection with 5498 features and 12 fields
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: 22.67231 ymin: -17.92831 xmax: 33.34831 ymax: -8.46428
## Geodetic CRS:  WGS 84
```

###Election data into districts' shapes

	ctr_n	ctr	yr	cst_n	cst	Rad_typ	MCC	MNC	Lct_A_C	CID	X0	Range	Chngb.1	Chngb.0	Created	Updated	Avrg_sg	geometry
1	Zambia	894	2006	Mfuwe	94	GSM	645	1	10115	30982	0	7830	16	1	2016-01-16	1565600127	0	POLYGON ((32.2369[illegible] -11.2491..
1.1	Zambia	894	2006	Mfuwe	94	GSM	645	2	520	6242	0	6217	3	1	2016-02-24	1507040057	0	POLYGON ((32.2369[illegible] -11.2491..
1.2	Zambia	894	2006	Mfuwe	94	UMTS	645	1	4108	232851	0	1000	1	1	2016-01-29	1454053584	0	POLYGON ((32.2369[illegible] -11.2491..
1.3	Zambia	894	2006	Mfuwe	94	UMTS	645	1	4108	231551	0	1672	3	1	2016-03-24	1481287116	0	POLYGON ((32.2369[illegible] -11.2491..
1.4	Zambia	894	2006	Mfuwe	94	GSM	645	1	10115	35262	0	14698	12	1	2016-07-03	1574074515	0	POLYGON ((32.2369[illegible] -11.2491..
1.5	Zambia	894	2006	Mfuwe	94	GSM	645	1	465	4051	0	3748	7	1	2016-03-25	1523586610	0	POLYGON ((32.2369[illegible] -11.2491..



Now I can rename variables knowing that :

CID = number of mobile cells

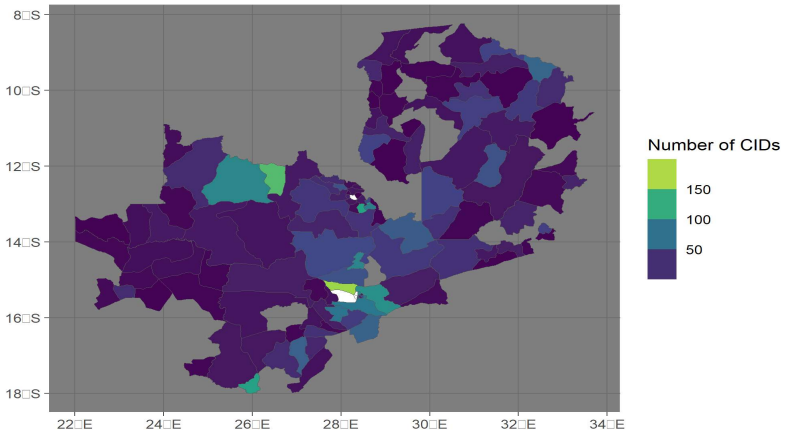
Geometry = location

Range = range of functioning of the tower

Statistical Analysis

First treatment assignation

I can now start to see how the number of a CID mobile internet cells is correlated with district level electoral outcomes from the lower chamber election in 2016. To do this I start by creating a new variables that is corresponding to the amount of internet mobile cells present in the electoral district, this is going to be the non-experimental treatment that I will try to leverage on in the analysis. After creating the new variable I plot it's geographical distribution using a white patch for variables with no data (NA):



Looking at the map above the lighter is the color the higher is the number of mobile cells and the picture suggests that there is heterogeneity in the intensity of the internet across different electoral districts.

At this point I can upload the data on the election for Zambia in 2016 taking them from https://electiondataarchive.org/data-and-documentation/clea-lower-chamber-elections-archive/ (https://electiondataarchive.org/data-and-documentation/clea-lower-chamber-elections-archive/) to see how it correlates with this first treatment assignation.

The data on the elections need to be cleaned, so I keep only useful column and then rename them to obtain the dataset below:

Lower Chamber 2016 election data

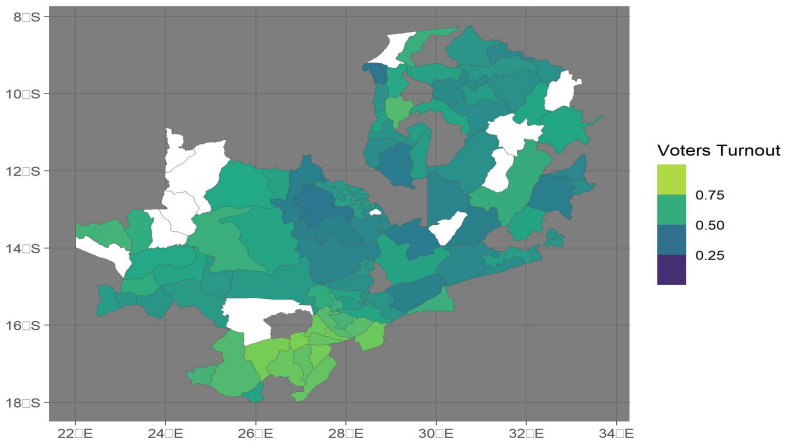
elec_district_name	party_name	party_code	n_eligible_voters	n_valid_votes	voter_turn	cvs1	n_votes	party_vpte_share	vote_cast	n_inval
BAHATI	Patriotic Front	23	43827	22222	0.5177858	0.8174782	18166	0.8174782	22693	
BAHATI	United Party for National Development	29	43827	22222	0.5177858	0.1825218	4056	0.1825218	22693	
BANGWEULU	Rainbow Party	46	54391	28961	0.5511757	0.0078727	228	0.0078727	29979	
BANGWEULU	Patriotic Front	23	54391	28961	0.5511757	0.6363731	18430	0.6363731	29979	
BANGWEULU	United Party for National Development	29	54391	28961	0.5511757	0.0704396	2040	0.0704396	29979	
BANGWEULU	Independent	6032	54391	28961	0.5511757	0.0302476	876	0.0302476	29979	

The aim at this stage is to merge also election data with the previous mobile_district dataset, to do so I first rewrite the name of the districts in a way that can match the one of the previous dataset and the I join them, obtaining the data below.

Merged data with electoral outcomes and mobile data

elec_district_nameelec_district_codeRangeCIDRad_typdategeometrytreat1party_nameparty_coden_eligible_votersn_valid_votesvoter_turncvs1n_votes
It is now possible to look at the distributions of the two outcome variables which are :

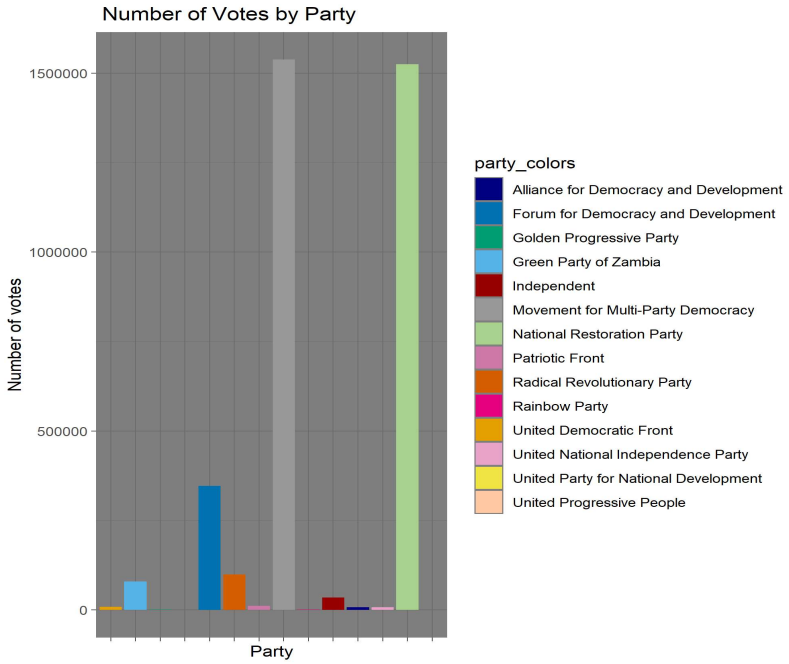
(I) Voters Turnout



It is already visible how the part of the country with a lower electoral participation in in the center of the country, and remember that white districts are those for which we have missing data.

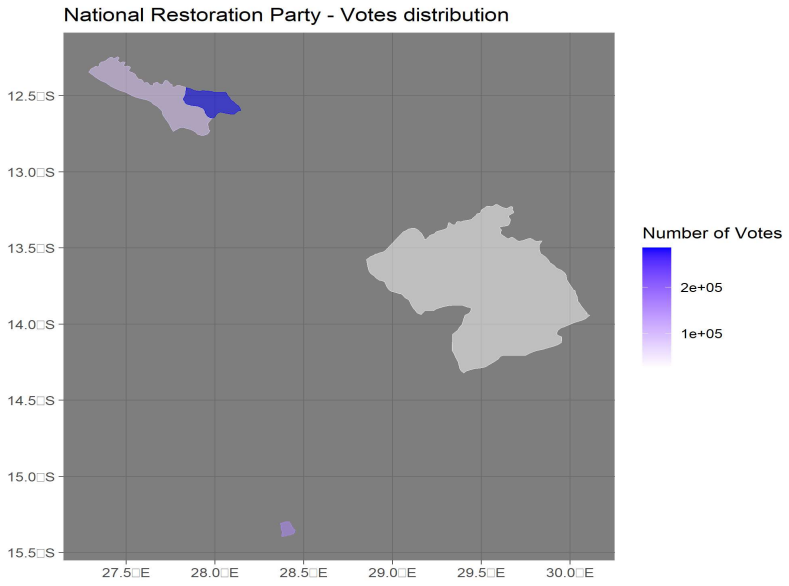
(II) Number of votes by party

To visualize this second outcome variable I first look at how the votes shares are distributed in all the country and then I detail district by district for the most important parties. Firstly we can see below a histogram with the vote share by party by creating a dataset with only party name and vote share and then plot it:

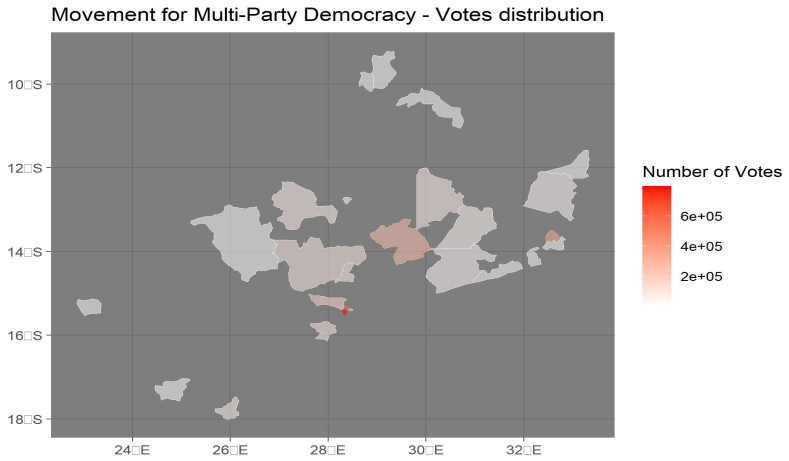


The two main parties seem to be National Restoration Party and Movement for Multi-Party Democracy.

Looking at how the votes for the two main parties are geographically distributed :

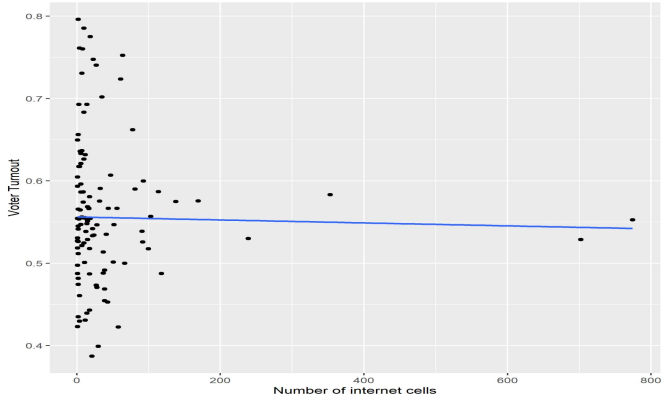


NRP is receiving more votes in the north-east part of the country while MPD is getting more support in the center where we could observe above the turnout is higher and there is relatively more internet access



Now I run the correlation between having at least one CID in your district and the voters turnout using a maximum likelihood estimator and I plot it filtering out NA:

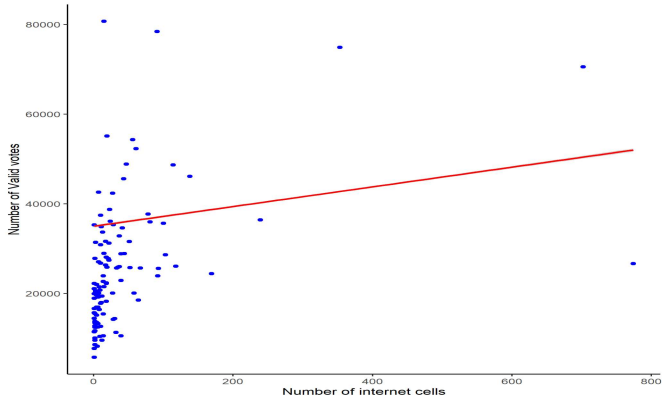
Correlation between internet access and voters' turnout



It looks like there is no correlation between the intensity of the treatment, internet access, and the voters turnout.

Now let's look at the correlation with the number of valid votes, maybe internet connection comes with more awareness on electoral competition rules:

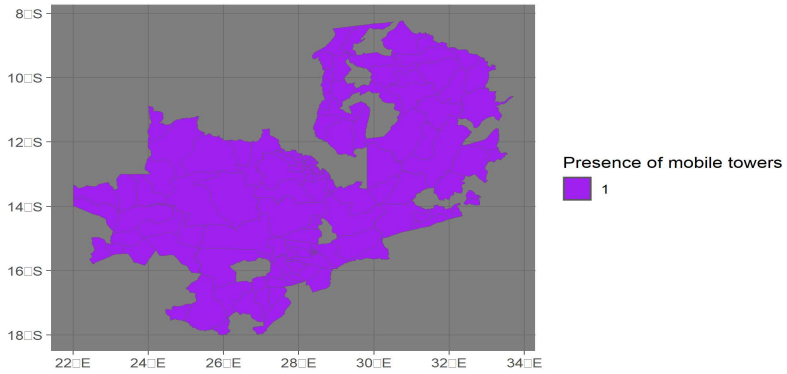
Correlation between internet access and number of valid votes



The picture here looks very similar, little to no correlation is shown.

Second Treatment Assignment

I can now assign the presence of internet access using the presence of treatment instead of the intensity of the treatment and see if the lack fo correlation is robust to this alternative treatment specification. To do this I create a dummy that is 1 if there is at least 1 CID in the electoral district and 0 otherwise adn then I plot it's distribution to see if there is heterogeneity in how this new way of assigning treatment look like.

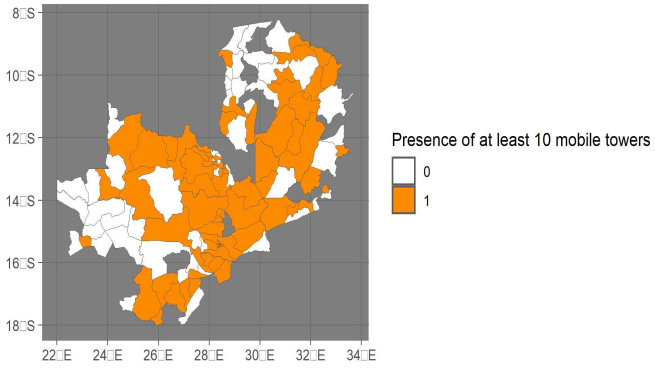


It is visible how in each district by 2016 there is still at least one mobile tower registered, so this way of assigning treatment is not useful for the purpose of the analysis.

To see how the intensity is changing I choose 3 different thesholds to try to identify the areas where there are more mobile cells registered.

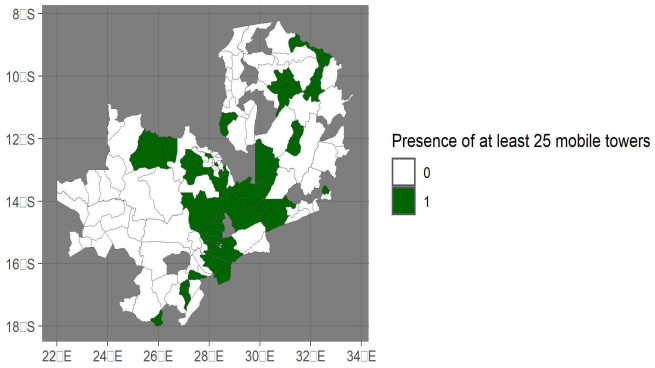
So I start by creating a treatment 2.1 where I assign to this new dummy the value of 1 if in the electoral district there are at least 10 mobile cells and then I create a treatment 2.2 with having at least 25 mobile cells in the district and at the end a treatment 2.3 for having more at least 50 cells.

Districts having at least 10 mobile cells



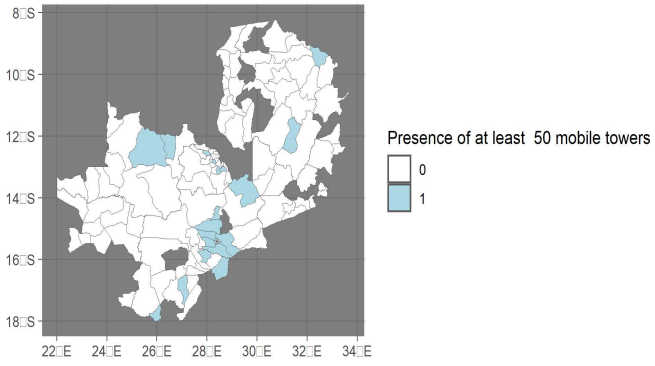
It is visible from the lowest threshold how the internet connection cells, as well as the most densly populated areas are located at the center of the country.

Districts having at least 25 mobile cells



When going over 25 only a bit more than the 30% of the country is colored and it is all located in nearby areas, indicating a spillover effect of the internet access most likely linked to the benefit of scale economy in technological infrastructures.

Districts having at least 50 mobile cells



After 50 cells the center is no more colored and around 10 districts located in the west-center of the country are remaining, indicating a correlation with other characteristics missing in the analysis regarding the socio-economic features of the districts' population but also that in 2016 the internet access was still not evenly distributed across different geographical areas.

Conclusion

In conclusion there is a lot of heterogeneity in terms of geographical location for both the turnout, the party vote share, the number od eligible votes and the presence of internet access mobile towers, but the patterns do not seem to be correlated after this first exploration.

Explained chunk of code

I choose to explain in detail this chunk of code :

```
vote_share <- election_2016 |>
  group_by(party_name) |>
  summarize(total_votes = sum(n_votes, na.rm = TRUE))
party_colors <- c("#E69F00", "#56B4E9", "#009E73", "#F0E442",
                          "#0072B2", "#D55E00", "#CC79A7",
                          "#999999", "#E6007E", "#980000",
                          "#000080", "#E9A3C9", "#A9D18E",
                          "#FFC8A3")
ggplot(vote_share, aes(x = party_name, y = total_votes, fill = party_name)) +
  geom_bar(stat = "identity") +
  labs(title = " Number of Votes by Party", x = "Party", y = "Number of votes") +
  scale_fill_identity(guide = "legend", labels = vote_share$party_name) +
  theme_dark() +
  theme(axis.text.x = element_blank())
```

1

I create a subset of the election 2016 data where I have the sum of the number of votes by party, to do this i group the n_votes based on the party name using the election_2016 data and I aggregate by summing up the n_votes.

2

I create a list of colors, one for each party that I have in the election data generating manually color scale.

3

I use ggplot setting the filtered votes by party as data, explaining in the aesthetic that I want the x axis to be filled with the party names, the y axis with how many votes and the fill should be based on the color scale i created manually above.

4

The stat = “identity” argument tells ggplot that the y variable is already in the correct scale and should be used as it is.

5

I use scale_fill_identity to produce a legend with the name of the party and the assigned color at the side of the graph.

6

Lastly I set the theme to be dark because I like the style and I set off the names of the parties in the x axis that was too crowded.